2007

On-Chip Digital Decoupling Capacitance Methodology

Eric Scott Bohannon

On-Chip Digital Decoupling Capacitance Methodology

by

Eric Scott Bohannon

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Science in Computer Engineering

Supervised by

Dr. P.R. Mukund
Department of Electrical Engineering
Kate Gleason College of Engineering
Rochester Institute of Technology
Rochester, NY
February 2007

Approved By:

Dr. P.R. Mukund
Primary Advisor – R.I.T. Dept. of Electrical Engineering

Dr. K. Hsu
Secondary Advisor – R.I.T. Dept. of Computer Engineering

Dr. D. Kudithipudi
Secondary Advisor – R.I.T. Dept. of Computer Engineering

Thesis Release Permission Form

Rochester Institute of Technology
Kate Gleason College of Engineering

Title: On-Chip Digital Decoupling Capacitance Methodology

I, Eric Scott Bohannon, hereby grant permission to the Wallace Memorial Library to reproduce my thesis in whole or part.

Eric Bohannon

Name

Date

Thesis/Dissertation Author Permission Statement

Title of thesis or dissertation: On-Chip Digital Decoupling Capacitance Methodology

Name of author: Eric Bohannon
Degree: masters
Program: EECB
College: Engineering

I understand that I must submit a print copy of my thesis or dissertation to the RIT Archives, per current RIT guidelines for the completion of my degree. I hereby grant to the Rochester Institute of Technology and its agents the non-exclusive license to archive and make accessible my thesis or dissertation in whole or in part in all forms of media in perpetuity. I retain all other ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Print Reproduction Permission Granted:

I, Eric Bohannon, hereby **grant permission** to the Rochester Institute Technology to reproduce my print thesis or dissertation in whole or in part. Any reproduction will not be for commercial use or profit.

Signature of Author: Eric Bohannon Date: _____

Print Reproduction Permission Denied:

I, _____, hereby **deny permission** to the RIT Library of the Rochester Institute of Technology to reproduce my print thesis or dissertation in whole or in part.

Signature of Author: _____ Date: _____

Inclusion in the RIT Digital Media Library Electronic Thesis & Dissertation (ETD) Archive

I, Eric Bohannon, additionally grant to the Rochester Institute of Technology Digital Media Library (RIT DML) the non-exclusive license to archive and provide electronic access to my thesis or dissertation in whole or in part in all forms of media in perpetuity.

I understand that my work, in addition to its bibliographic record and abstract, will be available to the world-wide community of scholars and researchers through the RIT DML. I retain all other ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation. I am aware that the Rochester Institute of Technology does not require registration of copyright for ETDs.

I hereby certify that, if appropriate, I have obtained and attached written permission statements from the owners of each third party copyrighted matter to be included in my thesis or dissertation. I certify that the version I submitted is the same as that approved by my committee.

Signature of Author: Eric Bohannon Date: _____

Dedication

I dedicate this work to my Sister. As time passes, I hope our relationship can become stronger.

Acknowledgements

I would like to thank my advisor, Dr. Mukund, for his guidance. Without him, this work would not have been possible. He is a mentor to me, both professionally and personally. I would also like to thank my committee members, Dr. Hsu and Dr. Kudithipudi. They were supportive of me throughout. To Mark Pude, thanks for spending countless hours in the lab with me, I am grateful for all of your help. To Chris Urban, thanks for making me laugh and reminding me that finishing actually meant something. Sharmila, Tejasvi, and Priya, I appreciate your friendship and thank you for welcoming me into the lab. To my family and friends, I would like to thank all of you; I appreciate all of your support. This work is a culmination of everything I have learned in my life.

Abstract

Signal integrity has become a major problem in digital IC design. One cause of this problem is device scaling which results in a sharp reduction of supply voltage, creating stringent noise margin requirements to ensure functionality. Reductions in feature size also result in increased clock speeds leading to many different high frequency noise producing components. As on-chip area increases to allow for more computational capability, so does the amount of digital logic to be placed, magnifying the effects of noisy interconnect structures.

Supply noise, modeled as $\Delta V = L\,di/dt$, is caused by rapid current spikes during a rise or fall time. Decoupling capacitors often fill empty on-chip space for the purpose of limiting this noise. This work introduces a novel methodology that attempts to quantify and locate decoupling capacitors within a power distribution network. The bondwire attached on the periphery of the face of the die is taken to be the dominant source of inductance. It is shown that distributing capacitance closer to the switching elements is most effective at reducing supply noise.

A chip has been designed using TSMC 90 nm technology that implements the ideas presented in this work. Simulation results show that noise fluctuations are high enough such that random placement of decoupling capacitance is not effective for large digital structures. The amount of interconnect generated on-chip noise increases with area, resulting in the need for an optimal decoupling scheme. As scaling continues, supply voltages and noise margins will decrease, creating the need for a robust decoupling capacitance methodology.

Table of Contents

List of Figures

List of Tables

Chapter 1 Introduction

1.1. Motivation for This Work

As transistors scale, the problems created by on-chip electrical noise continue to be magnified. Scaling is motivated by a continuous march towards cost reduction through increased functionality [1]. Although it results in increased packaging densities and high frequency operation, it also produces a decrease in noise margins. In order to maintain high chip yields, and thus effectively reduce the cost of integrated circuits, the problems presented by these small noise margins must be addressed.

Year	Gate Length (nm)	Vdd (V)	Power (W)	Clock Frequency (GHz)
2006	28	0.9	180	6.8
2007	25	0.8	189	9.3
2008	23	0.8	198	11
2010	18	0.7	198	15
2012	14	0.7	198	20
2014	11	0.6	198	28
2016	9	0.5	198	40
2018	7	0.5	198	53

Figure 1.1: 2005 NTRS [2]

The 2005 National Technology Roadmap for Semiconductors (NTRS) predicts transistors with gate lengths of 18nm and supply voltages of 0.7 V by 2010 [2]. These values are expected to reduce even further by the year 2018, as gate lengths approach 7 nm and supply voltages decrease to 0.5 V. Predictions are shown in Figure 1. It is obvious from Figure 1 that as supply voltages continue to drop, the amount of tolerable

noise will as well. Controlling noise on the shrinking power supply is essential to ensuring the functionality of digital circuits.

1.2. Importance of Voltage Scaling

It is important to understand why voltage scaling occurs and the impacts it has on device performance [3]. Consider the simple NMOS transistor shown in Figure 2.

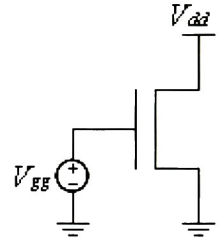

Figure 1.2: NMOS transistor

The on resistance of this transistor, R_{ON}, is modeled in the triode region as:

$$R_{ON} = \frac{V_{DS}}{I_{DS}} \tag{1.1}$$

$$I_{DS} = \beta(V_{GS} - V_{TN})V_{DS} \tag{1.2}$$

$$R_{ON} = \frac{1}{\beta(V_{GS} - V_{TN})} = \frac{1}{\beta(V_{GG} - V_{TN})} \tag{1.3}$$

If the transistor is scaled by a factor, f, the parameter β is altered via the gate oxide thickness (t_{OX}) [3]:

$$\beta = K_N' \frac{W}{L} \tag{1.4}$$

$$K_N' = \mu_N C_{OX} \tag{1.5}$$

$$C_{OX} = \frac{\varepsilon_{OX}}{t_{OX}} \tag{1.6}$$

$$t_{OX\,scaled} = \frac{t_{OX}}{f} \tag{1.7}$$

The dependence of the scaled oxide thickness ($t_{OX\,scaled}$) can be related back to the original β [3]. This is shown in the following equations.

$$C_{OX\,scaled} = \frac{\varepsilon_{OX}}{t_{OX\,scaled}} = \frac{\varepsilon_{OX}}{\left(\frac{t_{OX}}{f}\right)} = f\,C_{OX}' \tag{1.8}$$

$$K_N'_{scaled} = \mu_N C_{OX\,scaled} = \mu_N f\,C_{OX} = f\,K_N \tag{1.9}$$

$$\beta_{scaled} = K_N'_{scaled}\frac{W}{L} = f\,K_N'\frac{W}{L} = f\beta \tag{1.10}$$

Solving for $R_{ON\,scaled}$ without scaling any of its associated voltages yields:

$$R_{ON\,scaled\_NV} = \frac{1}{\beta_{scaled}(V_{GS} - V_{TN})} \tag{1.11}$$

$$R_{ON\,scaled\_NV} = \frac{1}{f\beta(V_{GS} - V_{TN})} \tag{1.12}$$

$$R_{ON\,scaled\_NV} = \frac{1}{f\beta(V_{GG} - V_{TN})} = \frac{R_{ON}}{f} \tag{1.13}$$

In order to create a baseline resistance as transistors are scaled and to physically ensure the oxide does not get damaged, the supply and threshold voltages are also scaled.

$$R_{ON\,scaled\_V} = \frac{1}{f\beta(\frac{V_{GS}}{f} - \frac{V_{TN}}{f})} \tag{1.14}$$

$$R_{ON\,scaled\_V} = \frac{1}{f\beta(\frac{V_{GG}}{f} - \frac{V_{TN}}{f})} = R_{ON} \tag{1.15}$$

Equation (15) shows that the scaled resistance is equal to the original resistance due to the effects of voltage scaling. These voltage reductions combined with smaller parasitic capacitances of the physical device itself allow for high speed switching. If voltage scaling did not occur, the resistance would be reduced causing an increase in power consumption.

$$P = V_{DS} I_{DS} = \frac{V_{DS}^{\ 2}}{R_{ON}} \tag{1.16}$$

$$P_{scaled\_NV} = \frac{V_{DS}^{\ 2}}{R_{ON\,scaled\_NV}} = \frac{V_{DS}^{\ 2} f}{R_{ON}} = V_{DS} I_{DS} f \tag{1.17}$$

Equation (17) shows an increase in power consumption, by the scale factor f, if device voltages are not scaled. With scaling, the amount of power consumed decreases:

$$P = V_{DS} I_{DS} = \frac{V_{DS}^{\ 2}}{R_{ON}} \tag{1.18}$$

$$V_{DS\,scaled} = \frac{V_{DS}}{f} \tag{1.19}$$

$$P_{scaled} = \frac{V_{DS\,scaled}^{\ 2}}{R_{ON\,scaled\_V}} = \frac{V_{DS}^{\ 2}}{f^2 R_{ON}} = \frac{V_{DS} I_{DS}}{f^2} \tag{1.20}$$

Equation (20) shows a decrease in power consumption due to scaling. This happened because the resistance remained constant and the voltage between the drain and source became smaller.

Voltage scaling is done not only to ensure physical functionality but to also take advantage of the benefits offered by scaling: decreased power consumption and area reduction.

1.3. Power Supply Noise

Decreases in supply voltage result in smaller noise margins, which, if violated can cause incorrect functionality. Supply noise occurs due to the parasitic resistive and inductive elements of the on-chip interconnect structure. Interconnect structures are scaled, but not as dramatically as transistors, in hopes of avoiding increased resistance, inductance, and possible crosstalk. Shrinking interconnect still results in less conductivity, allowing these parasitic elements to have more of an impact on the supply voltage. Interconnect resistance is an obvious source capable of degrading the power supply. On-chip inductances are of concern now because of device scaling and high frequency operation.

Figure 3 shows a simple schematic with an ideal supply and a realistic supply. The passive elements shown between the power and ground terminals in the realistic supply are responsible for generating noise. These elements are often ignored when analyzing digital circuits producing a supply like that shown in the ideal schematic.

The reactive elements (inductance) of interconnect are receiving more attention because of increased operating frequency. Modeled simply as $\Delta V = L\,di/dt$, supply noise is susceptible to high operating frequencies because of quick current spikes that occur when transistors switch in digital logic. High package densities contain a large number of transistors capable of drawing high amounts of current during a rise or fall time. If

each transistor is switched on the same clock edge, the amount of current drawn would be at a maximum and could possibly produce an intolerable amount of supply noise.

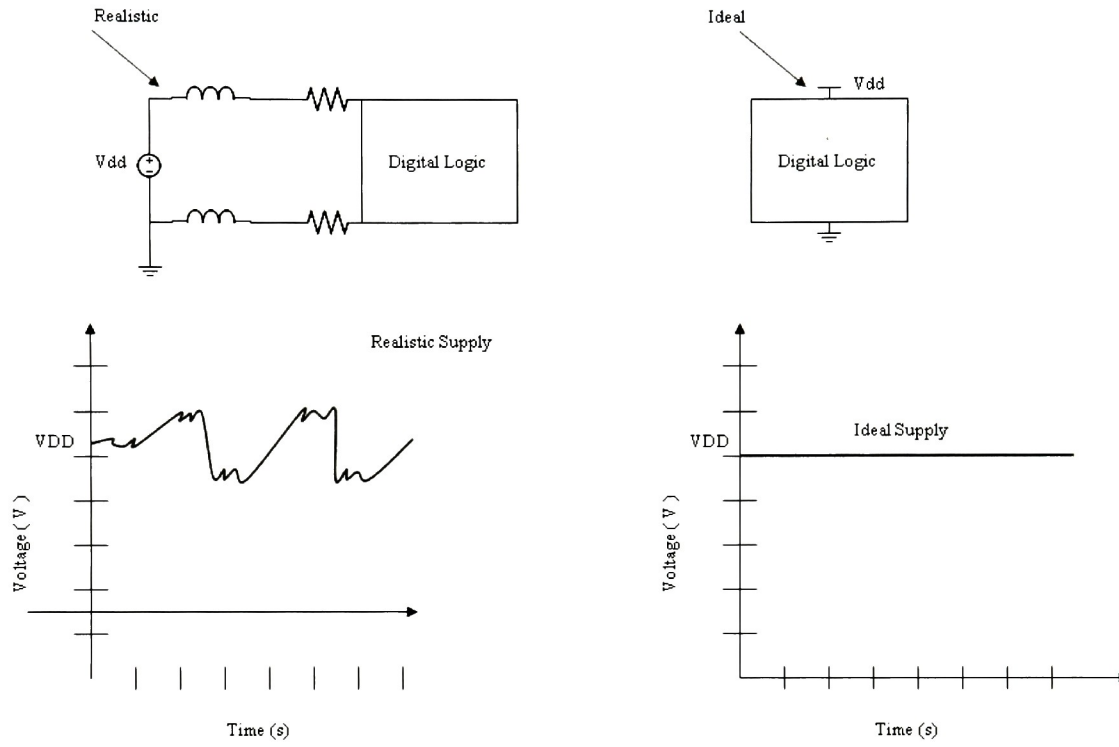

Figure 1.3: Ideal Power Supply vs. Actual Power Supply

The dominant inductive portion of supply noise comes from the bondwire that is attached on the periphery of the face of the die. The bondwire is used as interconnect between a chip and package. Inductance values for a 1 mm diameter bondwire range from 2-10nH depending on its length [4]. For example, a 5 mA spike through a 5 nH inductor that occurs every nanosecond results in a 25 mV fluctuation.

1.4. Decoupling Capacitance

Decoupling capacitance is often used to limit the amount of noise seen on a supply. These capacitors are used to decrease current bursts (δi) provided by the power supply that occur during a rise or fall time (δt). They act as temporary energy storage and are

placed in areas within a digital circuit to provide relief to switching transistors. Connected between the supply line and ground, this allows the capacitor to provide charge to the logic through a less impeded path. Figure 4 illustrates how a decoupling capacitor would be inserted into an interconnect structure.

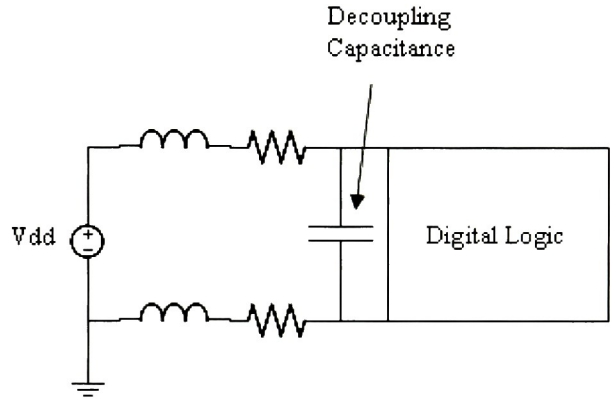

Figure 1.4: Decoupling Capacitance Connections

Effectively, decoupling capacitance reduces the amount of line inductance and resistance seen during high speed switching. Current decoupling capacitance strategies include placing fixed valued capacitors at random on-chip locations. For example, a designer may only have two different decoupling capacitance values for an entire design. Also, empty or extra space may be haphazardly filled with these capacitors [5]. While this scheme is effective, it is by no means efficient. If designers knew how to quantify and place decoupling capacitance, extra on-chip real estate could be saved or used to add functionality. This leads to a desire for an efficient decoupling capacitance scheme capable of meeting the strict noise margins introduced by scaling.

1.5. Other Noise Sources

Noise margin violations can occur from sources other than the power supply. Shot noise, flicker noise, and thermal noise are other sources that can cause circuit

malfunctions. A circuit should still be able to perform to specifications with all of these sources present. Because these other noise types exist, an activity factor, λ ,is defined that represents the amount of noise margin contributable fluctuations allowed on a power supply. Typical values range from 5% - 10% [6]. On a 1 V supply with an activity factor of 5%, the allowed amount of voltage fluctuations is 50 mV. This means that the supply should only go as low as 950 mV due to supply noise. The relationship is simply the supply voltage multiplied by the activity factor:

$$V_{fluc} = V_{dd}\lambda \tag{1.21}$$

An activity factor of 5% on a 5 V supply results in 250 mV of allowed fluctuations. This is a much easier specification to meet compared to the 1 V case. As voltages continue to scale and activity factors remain constant, the allowed amounts of fluctuations reduce.

The amount of noise present in a circuit at any given moment in time is unpredictable causing the desire for a large noise margin capable of handling multiple types of noise simultaneously. Although this work focuses on the reduction of power supply noise, it should be noted that multiple types of noise sources exist. Each different type of noise should have its own activity factor requirement. If a circuit is going to function correctly, the designer must take each of these noise sources into account.

Chapter 2 Literature Review

Previous work involving decoupling capacitance as a solution to power supply noise can be broken down into three domains. The first is how to represent the on-chip interconnect structure and its switching elements. The second involves the actual calculation of how much decoupling capacitance to use and where to place it. The third and final domain focuses on the need to avoid supply oscillations. The following sub-chapters break down each of these different domains.

2.1. Power Distribution Networks

On-chip interconnect structures, called power distribution networks, model the parasitic elements involved in power routing. The resistors and inductors that make up these networks contribute to power supply noise, which if large enough, may cause incorrect functionality in digital chips [7]. Natural decoupling capacitances exist in these networks but are often too small to be effective in reducing supply noise [8].

A methodology for approximating RLC tree structures as second-order systems is presented in [9]-[11]. The methodology extracts second order parameters like damping factor and natural frequency for symmetrical or asymmetrical RLC trees. These can then be used to calculate useful second order specifications like delay time, rise time, percent overshoot, and settling time.

This methodology can be used to compare different interconnect structures that have different decoupling capacitance strategies. For example, if an RLC tree structure is used to represent on-chip metal interconnect like in [12]-[16], decoupling capacitance can be inserted into the structure using different approaches. Each approach can then be

compared using the approximations in [9]-[11]. In [17], this approximation was extended to include a third initial moment, allowing for more accurate rise time and delay time calculations.

The delay time specification is significant due to the fact the interconnect structure must be treated as an RLC structure instead of an RC structure to accurately design for power supply noise. Previously, interconnect structures were modeled as RC networks where line delays were calculated using the Equivalent Elmore RC time constants [9], [18]. The introduction of inductance creates a second order type system which changes how delay time must be calculated. In [9]-[11] the differences in delay time using the RC time constant method and the second order RLC approach were significant. This leads to a conclusion that if on-chip interconnect structures are going to be modeled accurately, inductance must be included [19].

In [12], a linear programming technique is presented to maximize the allocation of existing white space in a layout floor plan for the placement of decoupling capacitors. An RLC structure is used to model the power distribution network and current sources are used to model switching digital logic. This is similar to what is done in industry where the focus lies more on optimizing existing white space instead of finding an optimal location within the interconnect. In [12], the capacitance for each node in the distribution network was calculated using the amount of charge that flows along its line and the allowed amount of voltage fluctuations based on a given activity factor. For example, a periodic current pulse flowing through a certain interconnect line can be integrated to obtain a value for charge. This amount of charge can be related to the allowed amount of voltage fluctuations to solve for decoupling capacitance

($C = \frac{\Delta Q}{\Delta V}$). This approach is termed the "Greedy Solution" because the amount of charge drawn during a single period is assumed to be at a maximum. This creates the greatest possible value for decoupling capacitance, which could occupy a large amount of area in layout. Because the charge flowing along a single interconnect line cannot be perfectly predicted, this approach is iteratively modified to determine the minimum amount of decoupling capacitance needed to satisfy noise requirements. This solution may lead to inaccuracies because the current sources used to model switching activities and the distribution networks are approximations. It may be best to err on the side of caution and follow a worst case approximation when calculating the amount of decoupling capacitance needed.

The practice of inserting current sources into the interconnect structure to represent transistor switching activity is common [6], [12]-[14], and [20]-[22]. The amount of granularity can differ depending on whether the current source is modeling a single gate or a large digital block. For example, consider a processor that has millions of transistors. If each primitive gate (NAND, NOR, NOT, etc) were modeled as a single current source, the problem quickly becomes too complex to be analyzed. Grouping the primitive blocks into larger structures reduces the complexities and simplifies calculations within the interconnect model. The linear piecewise current model described in [23] is simplistic yet effective and can be used multiple times in an interconnect structure to represent various switching activities.

In [14], Wang and Sadowska present a novel multi-grid based technique for the problem of on-chip power supply network optimization. The power distribution network is modeled using an RLC structure with current sources to representing switching

activities. As in [12], the focus was more on area optimization of the interconnect structure instead of determining how to quantify and place decoupling capacitance.

In [15], Chen discusses interconnect and circuit modeling techniques for full-chip power supply noise analysis. An RLC mesh is used to model the power distribution network. Supply noise occurring on any specific interconnect line was modeled as $\Delta V_L = \sum L \Delta I / \Delta t$, where L is the effective wire inductance of that interconnect line, ΔI is the current change during transition, and Δt is the rise or fall time. Using the rise or fall time as a way to calculate possible inductive noise is ideal because transistor switching occurs during this time. Current surges that happen during these transitions cause the current per second slope to be at a maximum. The summation of these inductive noise elements represent how the supply voltage can degrade through multiple inductors before reaching the switching logic.

2.2. Decoupling Capacitance Strategies

In [21], Smith introduces simplistic calculations that can be used to determine how much decoupling capacitance should be placed on a single node within a distribution network. He relates the amount of charge drawn during a current burst to the amount of tolerable voltage fluctuations allowed on the supply line. This approach is similar to [12] and provides a useful and flexible way to quantify decoupling capacitance. Combining this type of technique with the ability to determine an optimal decoupling capacitance location could result in a stable interconnect structure capable of suppressing supply noise.

In [6], a design technique is presented that determines the value and placement of on-chip decoupling capacitors for reducing power supply noise. An RLC structure was used to model the power distribution network; current sources were used to approximate transistor switching activities. Each link of the RLC structure was assumed to be of the same length with identical electrical characteristics. This does not take into account the possibility of highly inductive or resistive nodes in the distribution network that could create a noise hot spot. Decoupling capacitance within the distribution network was calculated by relating the amount of current drawn to the clock frequency and activity factor.

The experiments in [6] showed that as the decoupling capacitance was distributed away from the supply and closer to switching activities, noise decreased. No methodology was introduced to describe how the decoupling capacitors were distributed. The authors used the intuition that as the decoupling capacitors were moved closer to the switching elements, the high frequency charge they supplied would encounter a less noisy path. No theoretical metrics were derived to prove why noise reduction occurred or even if there was an optimal level of distribution within the power network. The distribution network was not examined for second order parameters like those presented in [9]. These parameters (natural frequency, damping factor) could have a significant impact on performance.

Intuitively, the amount of charge provided by each decoupling capacitor should encounter the least amount of resistance and inductance possible, which should result in less noise. In [24], Fan et al present an approach for quantifying decoupling capacitor locations on PCB's. Like [6], the modeling used showed that local decoupling has

benefits over global decoupling. This was done by placing decoupling capacitance further away from the supply and thus deeper into the PCB and was cited as an effective way to avoid noise caused by parasitic resistance and inductance. The work in [6] also showed that resonant frequencies increased as decoupling capacitance moved closer to switching activities. This effect is desirable because it allows for greater frequencies of operation. The concept of different levels of decoupling capacitance distribution is important and could be used to determine comparison metrics and theoretical approaches as to how decoupling capacitance should be placed within a power distribution network.

In [5], the need for an optimal decoupling capacitance placement strategy is noted. The equivalent size and current techniques presented can be used to simplify complex digital circuitry into a single gate. The amount of charge the gate is capable supplying during a transition time can then be solved using equations in [5]. This amount of charge can then be related to an activity factor, allowing for a calculation of decoupling capacitance.

2.3. Avoiding Supply Oscillations

In [20], Chen et al compare two different types of on-chip decoupling capacitors to target noise and leakage reduction: thin-oxide and thick-oxide. The thin-oxide gate capacitor was also noted for its common use in [15]. In [20] it is claimed that the parasitic resistance associated with decoupling capacitors is actually beneficial because it introduces a dampening effect into the power distribution network.

It is important that the selected decoupling capacitance does not cause resonance, $f = 1/\left(2\pi\sqrt{LC}\right)$, near any on-chip operating frequencies. Ideally, with decoupling capacitance inserted into the distribution network, resonance peaks won't

occur near any operating frequency. As the amount of capacitance increases, in order to meet strict noise requirements, the resonant frequency decreases and may approach an operating frequency. One solution to increase the resonant frequency is to decrease the amount of capacitance used, causing more high frequency charge to be sent from the power supply thus increasing the amount of noise seen on a supply line. In [20], a resonant peak near the operating frequency is acceptable if the series resistance of the capacitors can be used to dampen its effect. In this work, thin-oxide capacitors were targeted to reduce noise hot spots while the thick oxide capacitors were placed in less noisy areas because they produce less leakage current.

In [25]-[26], Larsson explains the need to avoid resonance in on-chip metal interconnect structures. It is shown that damping, through resistance, must be achieved to stop oscillations. The proposed interconnect model fails to account for on-chip inductance, which leaves possible noise hot spots unaccounted for. Only bondwire inductance is accounted for and is targeted as the sole cause for frequency dependent noise generation. While bondwire inductance is a dominant source of noise, the effects of scaling have shown that interconnect inductance plays a vital role in power supply degradation. This leads to more capacitance needing to be placed within the interconnect itself. The location of on-chip inductance relative to switching transistors could have a direct correlation on where decoupling capacitance should be inserted.

In [27], Chun et al introduce a methodology for the placement of decoupling capacitors for gigahertz systems. They note if the power distribution network is poorly designed it can result in ground bounce, supply compression, and electromagnetic interference (EMI). A target impedance, similar to what was done in [21], is derived that

must be met by the power distribution network over broad frequency ranges in order to avoid these problems. Intuitively, placing multiple capacitors in parallel could be done to meet impedance specifications over broad frequency ranges. Because capacitors aren't ideal and contain equivalent series resistances and inductances this approach is not applicable [2], [22], [26], and [28]. Impedance spikes occur due to the inductive component of the capacitor. The work in [27] failed to introduce how much decoupling capacitance should be inserted at specific areas of the distribution network and seemed to rely on a random selection of where capacitance was placed. Test cases and results were geared more towards PCB's rather than on-chip solutions.

In [13], Popovich et al introduce an approach to decoupling capacitance placement based on the distance of the capacitor from the power supply. They define an effective radius from the power supply as to where a decoupling capacitor should be placed. A triangle current pulse is used to represent transistor switching activities and an RLC mesh is used to represent the power distribution network. A mathematical model is derived to determine how much capacitance should be placed at certain distances away from the supply within the mesh. Theoretically, the distance between the supply and capacitor shouldn't matter unless the delay time introduced by the capacitor is greater than the rate at which the switching activity occurs. Line delay has less to do with distance from the power supply and more to do with the parasitic components in its path [9].

In [22], a decoupling capacitance methodology for multi-voltage power distribution systems is presented. As in [27], a calculated impedance is targeted to be met at different frequency ranges to ensure circuit operation. The different values of

decoupling capacitance used are based off of two voltage supply types; high voltage and low voltage. The amount of capacitance is not theoretically solved; one type of capacitor is chosen for high voltage and another for low voltage. This may be sub-optimal in the sense that extra area may be consumed and lower resonant frequencies may occur.

Chapter 3 Derivation of Decoupling Quantification and Location

This chapter is broken into four main sub-chapters. The first sub-chapter introduces general terminology and the different types of decoupling strategies that will be referenced in the rest of this document. The second derives a methodology for calculating how much decoupling capacitance to place at any node within a power distribution network. The third derives metrics to compare the different types of distribution strategies introduced in the first sub-chapter. The last sub-chapter presents a *Matlab* based simulation showing how these derivations can be used.

3.1. Distribution Strategies

Power distribution networks contain a passive element impedance representation of a circuit. Using these networks to represent interconnect in a digital IC provides a basis for examining where noise hot spots will occur [2], [6], [9]-[12], [17], [22], and [25]. An example of a power distribution network is shown in Figure 1. The capacitances seen in this figure represent parasitic decoupling capacitors that are inherent in the interconnect structure but are often too small to provide any noticeable reduction in noise. The inductance directly after the supply is a representation of the bondwire, which is a dominant source of noise. The current sources placed at the end of the tree are not part of the interconnect itself but characterize the switching activity of digital logic blocks. They can be formed using equivalent size or equivalent current techniques [5]. Combining these current sources with the interconnect structure provides a representation of how much supply noise each digital block could encounter.

In [9], a second order approximation is presented that characterizes distribution networks as RLC trees. The approximation can be done on symmetrical or asymmetrical structures and provides more accuracy than the RC time constant. Using the derivations in [9], interconnect comparisons can be made. Once decoupling capacitance is inserted into the tree, the interconnect structure is changed leading to different second order parameters. Using these parameters, specifications like overshoot, rise time, settling time, peak time, and delay time can be extracted as a basis for comparison.

Definitions for distribution networks, which are interchangeably called RLC trees or interconnect structures, are defined here for future reference. The first is depth, which represents the number of levels in the RLC tree, including the parent node. In Figure 1, the tree contains a depth of three. The number of nodes contained in a symmetrical binary tree can be represented as:

$$\#Nodes = 2^{depth} - 1 \qquad \textbf{(3.1)}$$

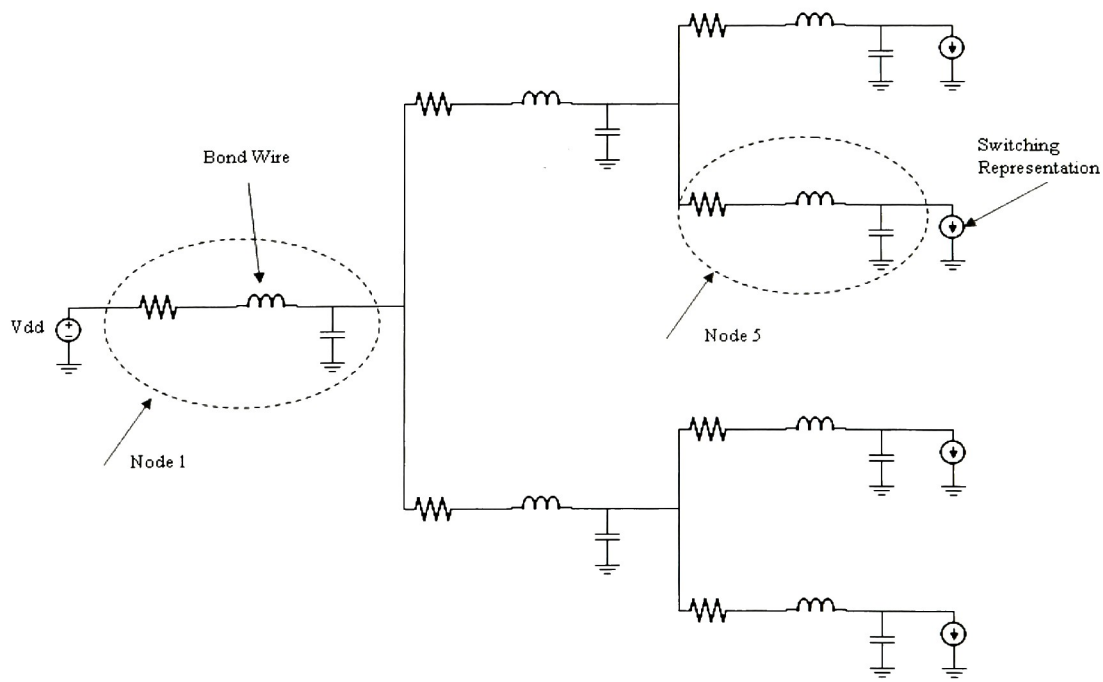

Figure 3.1: Power Distribution Network

Each RLC branch is identified as a node and assigned an index number. For example, in Figure 1, the first node (RLC branch) is the one directly after the supply and assigned index 1. The second node is the top RLC branch in the second level of the tree and assigned index 2. Index 3 is directly below the second. The fourth node starts at the top of the third level of the tree. The pattern continues as nodes 5, 6, and 7 are directly below the fourth node. When decoupling capacitance is placed at one of these nodes, it replaces the already existing capacitance.

Three types of general decoupling placement strategies are derived based on Figure 1. They are explained in the following three sub-chapters. Note in the following figures that all inherent parasitic capacitances have been removed. This is done to make the strategies simpler to understand and easier to visualize.

3.1.1 Lumped

The first strategy is lumped, which is defined as placing a single capacitor directly after the power supply and its associated impedances (bondwire and line resistance). This capacitor acts as temporary energy storage for the rest of the circuit and is placed on node 1 in Figure 2. It would be responsible for providing high frequency charge to all switching devices within the RLC tree. This capacitor only limits charge through the bondwire and its line resistance. Charge supplied by this capacitor encounters all forward parasitic elements in its path. Although it is generally accepted that this solution is impractical, it provides a basis for comparison and analysis.

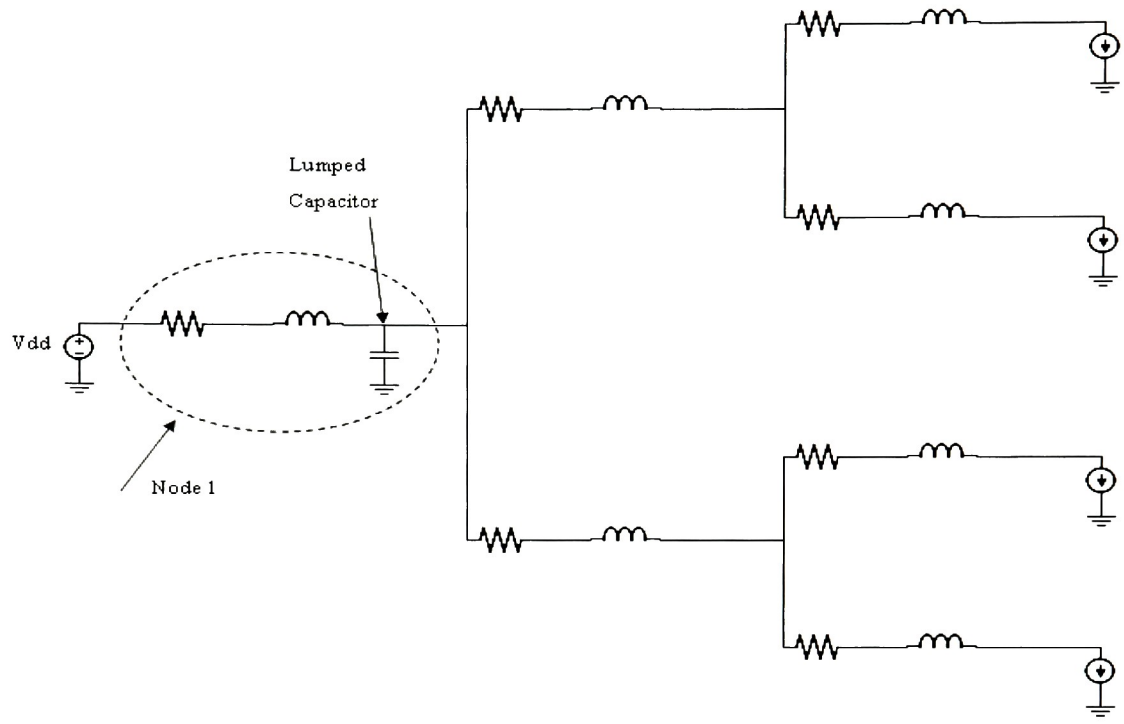

Figure 3.2: Lumped Decoupling Strategy

3.1.2 Distributed

The second strategy is distributed, which is defined as placing a single capacitor in front of each current source in the RLC tree. A current source could be represent a primitive gate (NOT, NAND, NOT, etc) or a group of primitive gates. The charge supplied by these capacitors does not encounter any parasitic elements on the RLC tree. Physically, this may not be realistic but it is assumed if a capacitor is placed directly in front a digital block, the impedance between them will be negligible. These capacitors act as energy storage for the current source they are in front of. An example is shown in Figure 3; capacitors are placed on nodes 4, 5, 6, and 7.

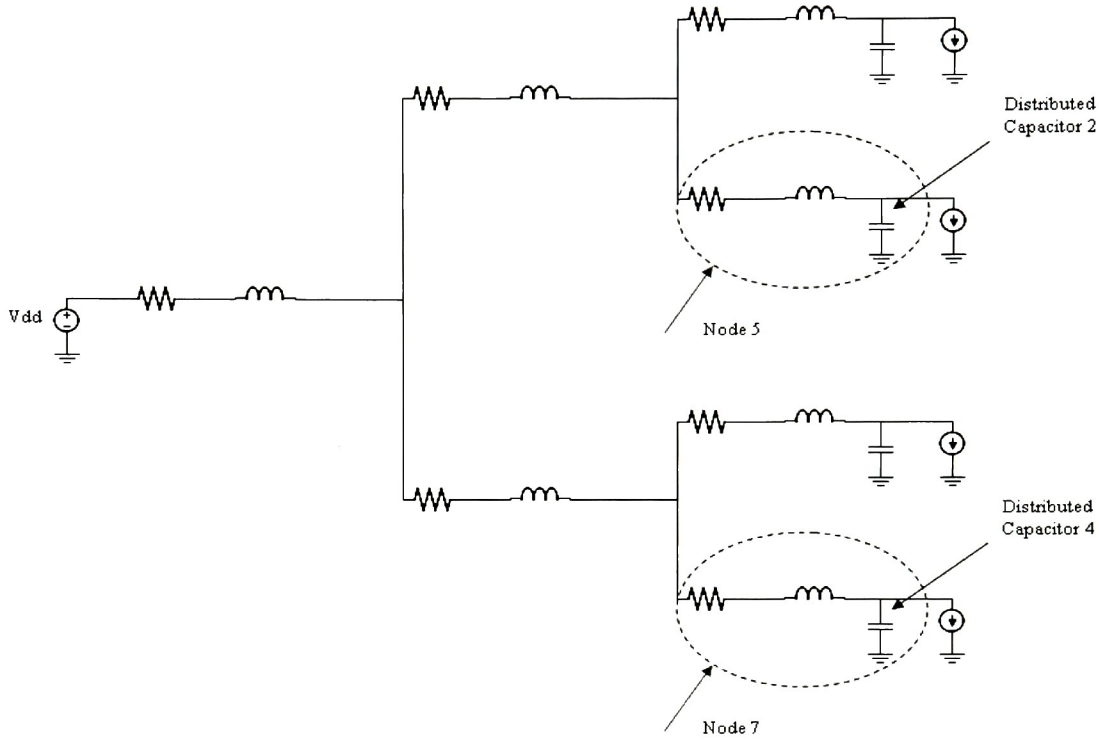

Figure 3.3: Distributed Decoupling Strategy

3.1.3 Grouped

The third strategy is grouped, which lies in-between lumped and distributed and is defined as placing capacitance in front of multiple cells in the circuit. Each value of capacitance acts as an energy storage for the group of cells it precedes and is shown by capacitors at nodes 2 and 3 in Figure 4. If the depth of the tree in Figure 4 were to increase, there would be more than one way to implement a grouped strategy. For example, if an RLC tree has a depth of 10 then placing a capacitor on level 1 is a lumped approach. Placing capacitors along level 10 would be distributed. Choosing a level between 2 and 9 and placing capacitors along that level is a grouped strategy. In general, for a tree with depth N, a grouped strategy would be followed by choosing a level between N+2 and N-1 and placing capacitors along it.

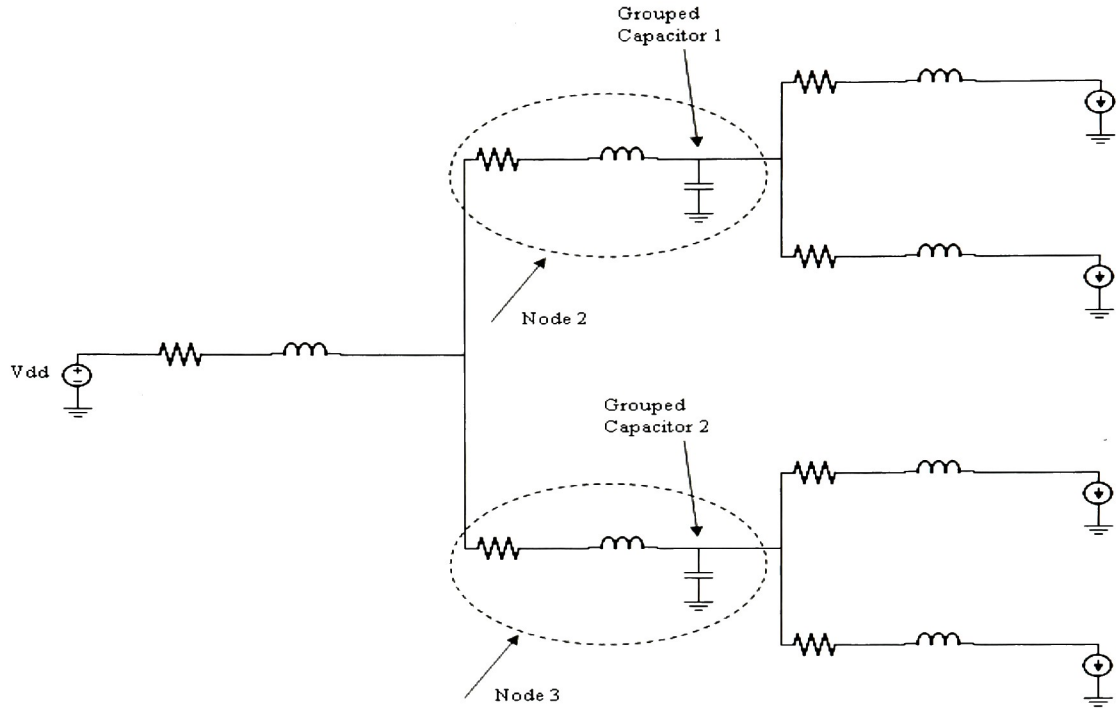

Figure 3.4: Grouped Decoupling Strategy

3.2. Capacitance Calculation

This sub-chapter encapsulates three more sub-chapters. The first explains the models used for the current sources in the RLC tree structures. The second introduces assumptions needed to calculate decoupling capacitance. The third derives the decoupling capacitance calculation.

3.2.1 Current Models

The current sources used to model transistor switching lend themselves nicely to CMOS technology because switching occurs during the rising or falling edge of a clock cycle. Also, a minimal amount of power is consumed when no switching is occurring, keeping the supply relatively stable. The amount of current each source draws cannot be exactly calculated because digital structures do not follow the same logic on every clock

cycle. Consider a clocked digital block, on one clock cycle the block may draw 2 mA of current and on the next clock cycle it may draw 5 mA of current. The amount of current the block draws is dependent on various factors such as inputs, outputs, and previous states. Because the amount of current drawn by a digital block changes, certain models must be used that accurately depict current consumption. One approach is to use a simulator such as *Cadence* to extract current characteristics. For example, a circuit schematic can be entered into software like *Cadence* and a transient analysis can be performed on it. Once the analysis is complete, current data can be extracted through a plot. An example is shown in Figure 5.

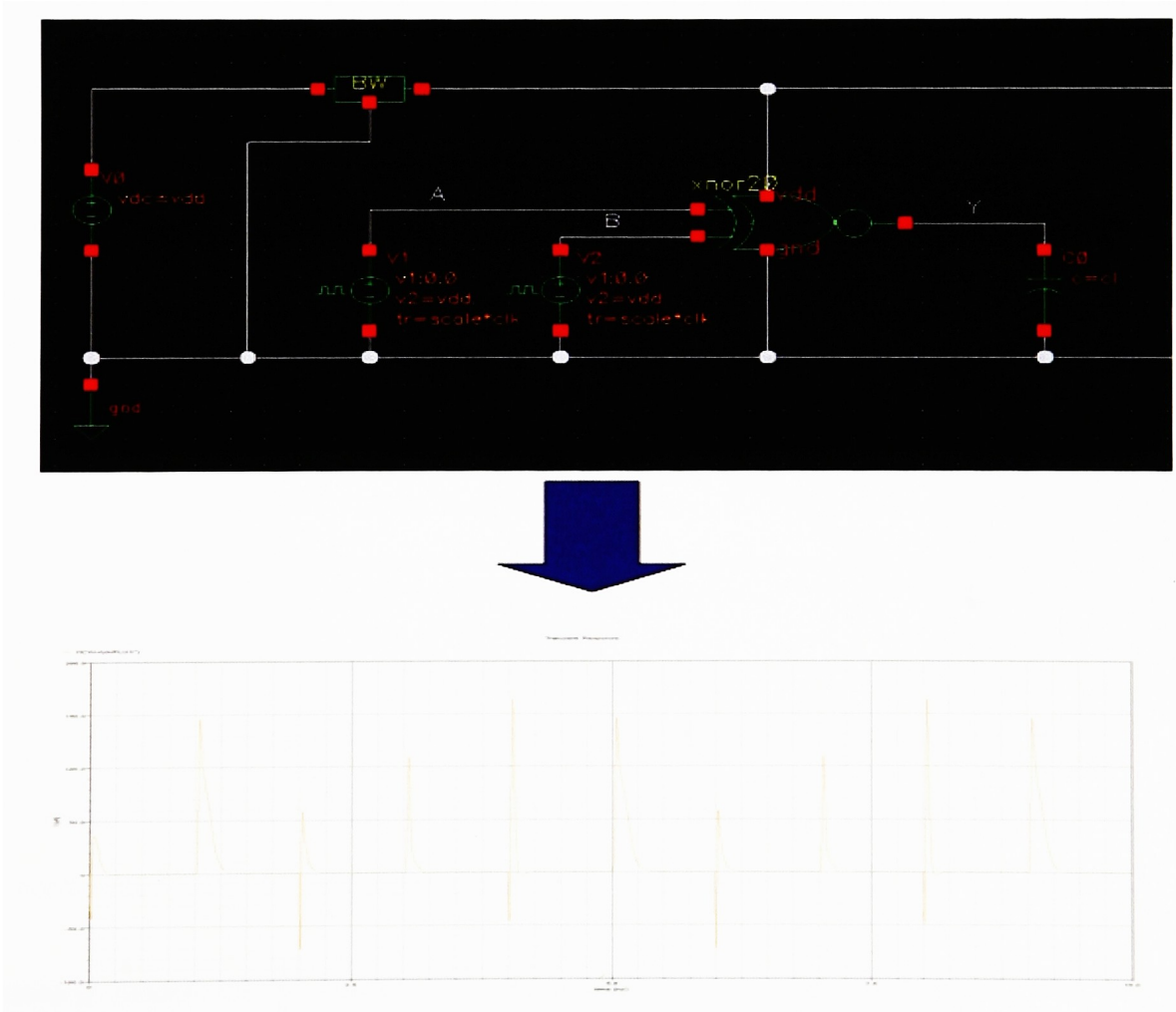

Figure 3.5: Schematic to Current Plot

Another approach is the complex gate collapse techniques (equivalent size devices and equivalent current devices) presented in [5]. These provide a mathematical analysis to determine how much current a digital block is capable of sourcing or sinking. In [5], current models were specifically used that targeted short channel transistors. This is important because smaller transistors do not follow the square law approximations commonly taught in introductory electronics courses.

The following sub-chapter presents a simple model that will be used in future chapters to mathematically represent the current draw of digital logic.

3.2.1.1 Linear Piecewise

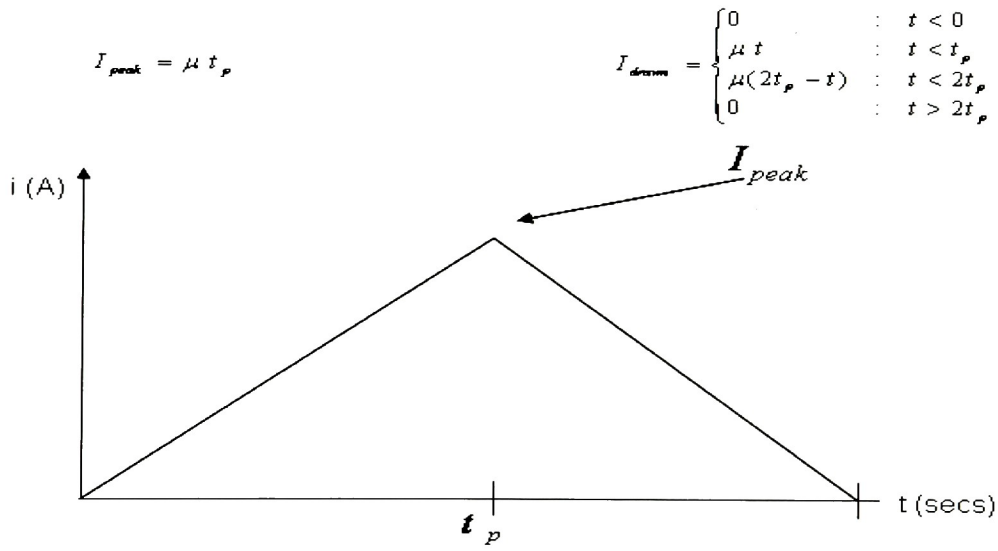

Figure 3.6: Current Approximation

The current sources in Figure 1 provide a theoretical basis for determining how much charge will need to be supplied by each decoupling capacitor. The model used in [13] represented them as triangular pulses with two parameters; μ and t_p. The parameter μ models the $\frac{\delta i}{\delta t}$ current bursts associated with a specific digital block. The

parameter t_p models the amount of time the digital block needs to reach its peak current. A linear piecewise function for this model is shown in Figure 6. The model creates a simple triangular current pulse that resembles what is seen when digital logic switches on a clock edge.

This current model can be applied to Figure 1. Each current source placed on the last level is assigned a unique μ and t_p. The source at node 4 has parameters μ_4 and t_{p4} while the source at node 5 has parameters μ_5 and t_{p5}. This pattern continues for nodes 6 and 7. The amount of charge to be drawn by these nodes can be gotten by integrating the current drawn during a switching transition.

$$\begin{aligned} Q_{node} &= \int_0^{2t_p} I_{drawn} = \frac{1}{2}bh \\ Q_{node} &= \frac{1}{2}2t_p I_{peak} \\ Q_{node} &= \frac{1}{2}2t_p \mu t_p = t_p{}^2 \mu \end{aligned} \qquad (3.2)$$

3.2.2 Distribution Network Assumptions

Once unique μ and t_p values have been assigned to the nodes on the last level of the tree, the preceding nodes must be assigned these parameters. In order to do this, the following assumptions are made:

1. The worst case $\frac{\delta i}{\delta t}$ slope at any node occurs when its forward nodes switch at the same time.
2. The amount of charge seen during a switching period is a summation of its forward nodes.

The first assumption relates switching in digital circuitry to worst case current draw. Because digital logic usually switches on a clock edge, it is safe to assume it will draw a maximal amount of current during this time. If all logic switches on the same clock edge, a worst case current slope occurs.

The second assumption relates the amount of charge switched by a parent node to its children nodes. For example, if a single node has three children, the amount of charge flowing through it will be a summation of the amount of charge flowing through its children.

Mathematically, these assumptions can be written to solve for μ and t_p at any node that isn't in the last level of the tree (where n is the number of children at the node being analyzed):

$$\mu_{new} = \sum_{i=1}^{n} \mu_i \tag{3.3}$$

$$Q_{new} = \sum_{i=1}^{n} Q_{node_i} = \sum_{i=1}^{n} t_{p_i}{}^2 \mu_i \tag{3.4}$$

$$\begin{aligned} t_{pnew}{}^2 \mu_{new} &= Q_{new} \\ t_{pnew} &= \sqrt{Q_{new} / \mu_{new}} \\ t_{pnew} &= \sqrt{\sum_{i=1}^{n} Q_{node_i} \Big/ \sum_{i=1}^{n} \mu_i} \\ t_{pnew} &= \sqrt{\sum_{i=1}^{n} t_{p_i}{}^2 \mu_i \Big/ \sum_{i=1}^{n} \mu_i} \end{aligned} \tag{3.5}$$

For example, in Figure 1 solving for these parameters at node 2 yields the following equations.

$$\mu_2 = \mu_4 + \mu_5$$
$$Q_2 = Q_4 + Q_5 = {t_{p4}}^2 \mu_5 + {t_{p5}}^2 \mu_5$$
$$t_{p2} = \sqrt{\frac{{t_{p4}}^2 \mu_5 + {t_{p5}}^2 \mu_5}{\mu_4 + \mu_5}}$$

This process can be repeated at node 1 making use of the results from nodes 2 and 3:

$$\mu_1 = \mu_2 + \mu_3$$
$$Q_1 = Q_2 + Q_3 = {t_{p2}}^2 \mu_2 + {t_{p3}}^2 \mu_3$$
$$t_{p1} = \sqrt{\frac{{t_{p2}}^2 \mu_2 + {t_{p3}}^2 \mu_3}{\mu_2 + \mu_3}}$$

3.2.3 Decoupling Capacitance Derivation

The parameter λ will be used as the activity factor (introduced in Chapter 2). It characterizes the amount of tolerable noise on the supply line. It is usually given as a percentage. For example, if a 1 V power supply is targeted to stay within 5% then the theoretical minimum supply voltage along any interconnect path would be 0.95 V. Generally, the minimum supply voltage can be given as:

$$V_{dd\,\min} = V_{dd} - V_{dd}\lambda = V_{dd}(1-\lambda) \qquad (3.6)$$

The amount of decoupling capacitance, C_{decap}, used to keep a node within this margin is shown in Equation (7). Where Q_{node} represents the amount of charge to be supplied by a specific node that contains decoupling capacitance and V_{fluc} is the amount of tolerable fluctuations.

$$C_{decap} = \frac{Q_{node}}{V_{fluc}} \tag{3.7}$$

The value for Q_{node} can be obtained from Equation (2) or from software simulation. If a software tool like *Cadence* is used, a circuit can be simulated to obtain a transient current plot (Figure 5). The current plot can then be integrated over a clock period to obtain a value for charge. This amount of charge is representative of what would need to be supplied by a decoupling capacitor to avoid power supply noise. In order to determine which clock period to integrate to obtain the charge value, multiple approaches can be used. One approach would be taking the worst case or maximum amount of charge seen in a single period. This could result in a large decoupling capacitor that consumes a lot of area. Another approach could be taking an average of the charges seen on each different clock edges.

Determining the right amount of charge to use is situation dependent. For example, if a system has strict noise requirements, a maximal amount of charge could be used. If a system has lenient noise requirements, a smaller amount of charge could be used.

Ideally, the amount of tolerable fluctuations at any node would be $V_{dd}\lambda$ but resistance losses that occur in the path between the supply and target node must be accounted for. V_{line} represents the voltage at any node due to DC resistance losses. This is shown in Figure 7 with resistors in parallel with each current source. When no switching is occurring, a low frequency current loop exists from the supply to ground. Ideally no current would be drawn when transistors aren't switching but in reality a

resistance exists between the drain and source of transistors (Rds_{off}) that causes voltage drops.

$$V_{line} = V_{dd} - \sum_i R_i I_i \qquad (3.8)$$

R_i = *amount of encountered series resistance at node*

I_i = *amount of DC current flowing at node*

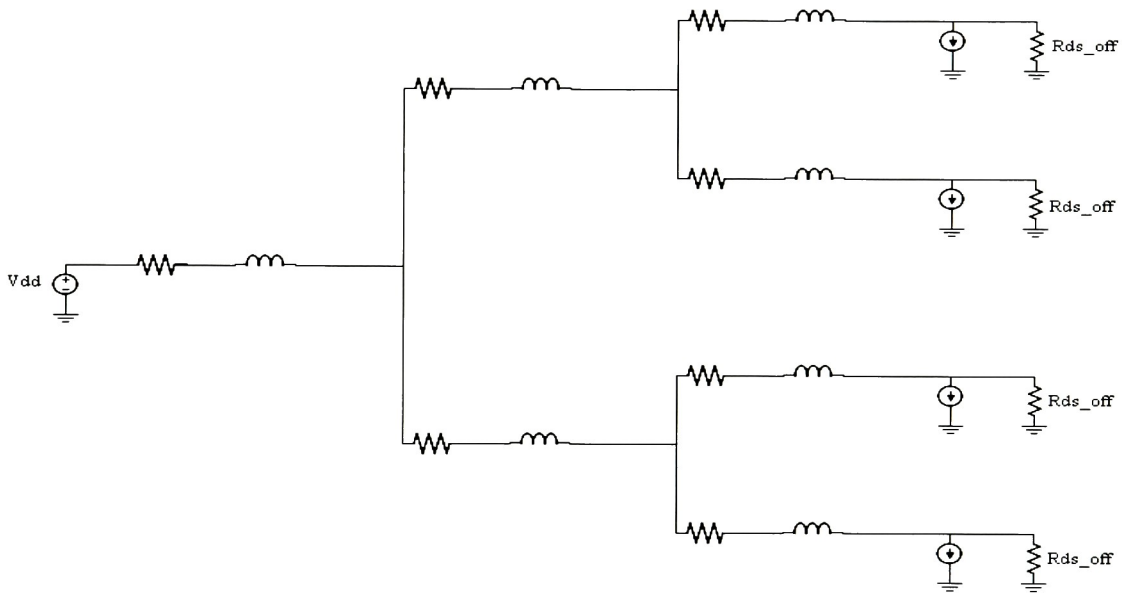

Figure 3.7: Off Resistance

For example, in Figure 5 to determine the V_{line} parameter for node 4, the resistance at nodes 1, 2, and 4 must be accounted for.

$$V_{line4} = V_{dd} - R_1 I_1 - R_2 I_2 - R_3 I_3 \qquad (3.9)$$

$$V_{line4} = V_{dd} - R_1 I_{dc1} - R_2 I_{dc2} - R_3 I_{dc3} \qquad (3.10)$$

Because DC line drops exist, a slight increase in decoupling capacitance occurs as placement moves from lumped to distributed. This is due to the fact that the original λ accounted for only V_{dd} and not any associated resistive interconnect drops. This leads

to the formation of λ_{node} which adjusts the original λ to take resistive interconnect drops into account. Each node within the tree has a unique λ_{node}.

$$\lambda_{node} = 1 - \frac{V_{dd\,\min}}{V_{line}} \tag{3.11}$$

Solving for V_{fluc} :

$$V_{fluc} = V_{dd}\lambda_{node} \tag{3.12}$$

$$V_{fluc} = V_{dd}\left(1 - \frac{V_{dd\,\min}}{V_{line}}\right) \tag{3.13}$$

$$V_{fluc} = V_{dd}\left(1 - \frac{V_{dd}(1-\lambda)}{V_{dd} - \sum_i R_i I_i}\right) \tag{3.14}$$

$$V_{fluc} = \frac{V_{dd}(V_{dd}\lambda - \sum_i R_i I_i)}{V_{dd} - \sum_i R_i I_i} \tag{3.15}$$

Substituting back into to (7):

$$C_{decap} = \frac{Q_{node}}{V_{fluc}}$$

$$C_{decap} = \frac{Q_{node}}{V_{dd}\lambda_{node}} \tag{3.16}$$

$$C_{decap} = \frac{Q_{node}}{V_{dd}\left(1 - \frac{V_{dd\,\min}}{V_{line}}\right)} \tag{3.17}$$

$$C_{decap} = \frac{\mu_{node}\, t_{p_{node}}{}^2}{V_{dd}\lambda_{node}} \tag{3.18}$$

$$C_{decap} = \frac{\mu_{node}\, t_{p_{node}}^{\,2}}{\left(\frac{V_{dd}(V_{dd}\lambda - \sum_i R_i I_i)}{V_{dd} - \sum_i R_i I_i} \right)} \tag{3.19}$$

$$C_{decap} = \frac{(V_{dd} - \sum_i R_i I_i)\mu_{node}\, t_{p_{node}}^{\,2}}{V_{dd}(V_{dd}\lambda - \sum_i R_i I_i)} \tag{3.20}$$

Equation (20) provides the ability to determine how much decoupling capacitance should be placed at any node within any generic interconnect structure. First, a distribution strategy is chosen. Then, each node within the chosen level of distribution is assigned a capacitor based on Equation (20). For example, if a grouped level 2 placement strategy was going to be followed in Figure 5, (20) would be applied to nodes 2 and 3.

3.3. Comparison Metrics

Placement metrics must be derived to guide how decoupling capacitance in a given RLC tree could be placed. The following derivations are done for a symmetrical binary tree with depth N. Two metrics are derived. The first metric is called inductive level noise and measures the average amount of noise seen at a particular level in an RLC tree. The second metric is called look-ahead inductive noise and measures the amount of noise seen along unique paths within an RLC tree.

3.3.1 Inductive Level Noise

The amount of inductive noise seen at any particular node is shown in Equation (21). This represents choosing a single node within the RLC tree and inserting

that nodes μ parameter(s) to determine the amount of noise the inductance at that node is capable of producing. If the node chosen is on the last level of the tree, there is only one μ parameter. If the node is not on the last level of the tree, then it has two children (because it's a binary tree) and utilizes two μ parameters. The μ parameters are summed to determine the amount of noise the inductance that node is capable of producing. The fact that the μ parameters are summed is one of the worst case assumptions presented in this chapter. The variable i is a counter that iterates over all nodes in the tree.

$$
\begin{aligned}
& i = 1 \ldots 2^n - 1 && \qquad (3.21) \\
& \Delta V_i = L_i(\mu_{2*i} + \mu_{(2*i)+1}) && 1 \le i < 2^{n-1} \\
& \Delta V_i = L_i \mu_i && 2^{n-1} \le i < 2^n - 1
\end{aligned}
$$

For example, if node 4 was selected from the RLC tree and had a μ parameter of 5 mA/ps and an inductance of 200 pH the amount of inductive noise capable of being produced is:

$$
\begin{aligned}
& \Delta V_4 = L_4 \mu_4 \\
& \Delta V_4 = 200\ pH * 5\ mA/ps = 1\ V
\end{aligned}
$$

Equation (21) can be manipulated to determine the average amount of noise seen at a particular level in the tree. For example, in Figure 1 the average amount of supply noise in level 2 would be the average of the inductive noise between nodes 2 and 3.

$$\Delta V_{AVG2} = (\Delta V_2 + \Delta V_3)/2$$

The importance of this measurement is that it can be used to show how much noise is possible of being generated at each level in the tree. Inductive noise will generally decrease as the analysis moves from the first level of the tree to the last level of

the tree. This occurs because the amount of current in the RLC tree decreases along each node as it moves from the supply to the current sources. Writing a general equation that applies to a binary tree; j is a variable that iterates over all tree levels:

$$j = 1 \ldots n$$

$$\Delta V_{AVGJ} = \left(\sum_{i=2^{j-1}}^{2^j - 1} V_i \right) \Big/ 2^{j-1} \tag{3.22}$$

If a sudden noise drop-off occurs at a certain level within the RLC tree, it could point to an optimal level of distribution. For example, if a 10 level RLC tree was being analyzed and the average amount of inductive noise decreased from 100 mV on level 5 to 15 mV on level 6, the optimal level of distribution may be level 5.

If there are outliers within the average inductive noise metric, they can be found by observing the maximum amount of noise produced on a single node in each level. The following equation calculates the maximum amount of inductive noise possible at a single node for each level within an RLC tree. The variable j iterates over all tree levels and the variable i iterates over each node within a specific tree level.

$$j = 1 \ldots n \tag{3.23}$$

$$i = 2^{j-1} \ldots 2^j - 1$$

$$\Delta V_{MAXJ} = \max(\Delta V_i)$$

This is useful to show a worst case noise situation. If the maximum amount of noise at a given node in a particular level is high, it could point to another possible noise hot spot within the RLC tree.

3.3.2 Look-Ahead Noise

It is important to analyze how the inductive noise propagates through the RLC tree. A certain path within the tree may be excessively noisy and identifying it may help

determine a decoupling capacitance solution. Look-ahead noise is defined as the summation of inductive noise along a unique path within an RLC tree. For example, in Figure 1 the amount of look-ahead noise that propagates from the supply to node 5 would be:

$$\Delta V_{LA5} = \Delta V_1 + \Delta V_2 + \Delta V_5$$

For a random node, x, within a binary RLC tree, the look-ahead noise can be calculated using the pseudo-code shown in Figure 8.

A noise matrix, N, is defined that holds all unique paths in the tree from the supply to any given current source. Each row represents one single path while each column contains the amount of inductive noise at a particular node in that path. For example, in Figure 1 there are 4 unique paths from the starting node to a current source. The noise generated by the bondwire is the dominant source of inductive noise. Each unique noise path in the RLC tree has the noise generated by the bondwire within it. Within the noise matrix N, the noise generated by the bondwire is neglected because it provides nothing more than a voltage offset to each path within the tree. For example, analyzing Figure 1, the look-ahead noise calculations including the bondwire are shown below for each of the 4 unique paths from the supply to the currents sources:

$$\Delta V_{LA4} = \Delta V_1 + \Delta V_2 + \Delta V_4$$
$$\Delta V_{LA5} = \Delta V_1 + \Delta V_2 + \Delta V_5$$
$$\Delta V_{LA6} = \Delta V_1 + \Delta V_3 + \Delta V_6$$
$$\Delta V_{LA7} = \Delta V_1 + \Delta V_3 + \Delta V_7$$

Notice that the common term in each of the four paths is ΔV_1, or the noise generated by the bondwire. In each path, this term adds nothing more than an offset and

does not help differentiate the noise differences in the other paths. In terms of Figure 1, the matrix N would be defined as:

$$N = \begin{bmatrix} \Delta V_2 & \Delta V_4 \\ \Delta V_2 & \Delta V_5 \\ \Delta V_3 & \Delta V_6 \\ \Delta V_3 & \Delta V_7 \end{bmatrix}$$

```
//the function calc_LA_Noise returns the amount of look ahead noise seen
//at a particular node of an RLC tree
LA_Noise = calc_LA_Noise(node startNode){

LA_Noise = 0; //amount of look ahead noise calculated
calcNode = startNode; //node used for calculating is first set equal to
                      //starting node

while(calcNode != root){ //once the top of the tree is reached, stop the loop

//add noise calculation for each node
LA_Noise = LA_Noise + (calcNode.inductance * calcNode.currentSlope);

calcNode = calcNode.getParentNode(); //new node to calculate with is parent
}

//add the contributions of the root node
LA_Noise = LA_Noise + (calcNode.inductance * calcNode.currentSlope);
}}
```

Figure 3.8: Look-Ahead Noise Pseudo-Code

Note the that each row is missing the inductive bondwire noise, ΔV_1. If this value were in the matrix, each path would contain it and nothing of value would be added to the analysis. The role of this metric is to help identify possible decoupling capacitance solutions within an RLC tree. It is assumed that at least a lumped decoupling strategy is followed, which would restrict the amount of high frequency charge flowing through the bondwire. Including the bondwire noise within this analysis would offer nothing beneficial in determining how to place decoupling capacitance.

More generally, for a tree with depth n the noise matrix is defined as:

$$N = \begin{bmatrix} \Delta V_2 & \cdots & \Delta V_{2^{n-1}} \\ \vdots & \ddots & \vdots \\ \Delta V_3 & \cdots & \Delta V_{2^n-1} \end{bmatrix} \qquad (3.24)$$

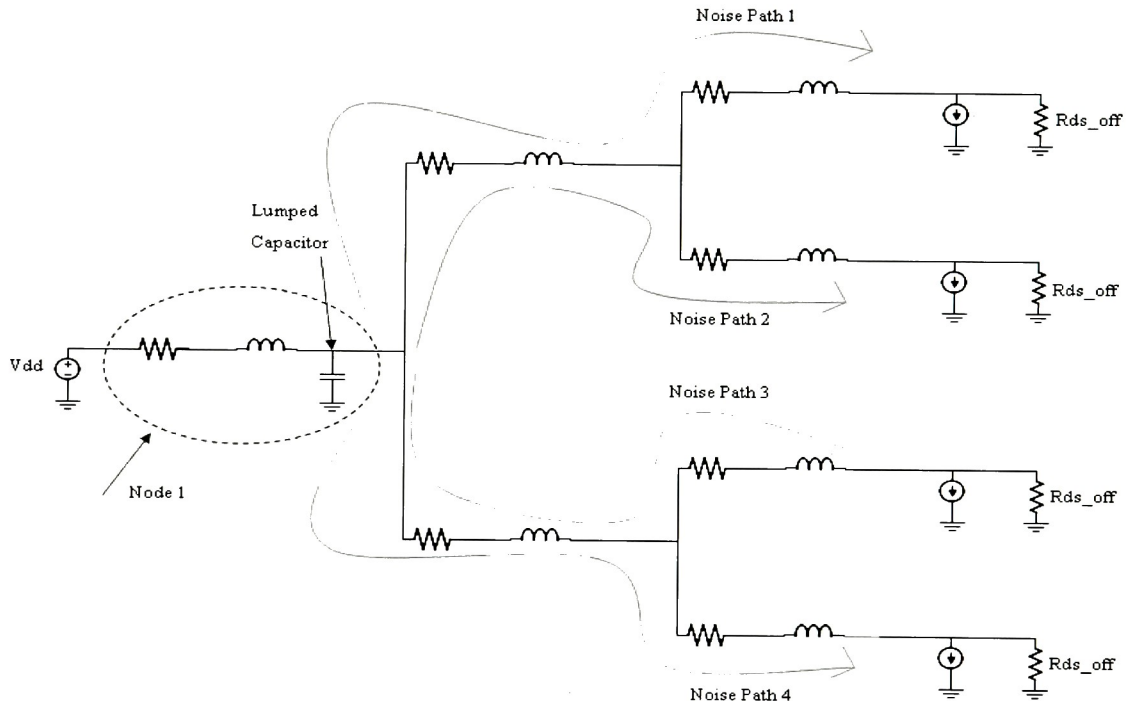

Figure 3.9: Lumped Capacitor Noise Paths

If a lumped strategy were followed in Figure 1 the majority of inductive noise would occur in nodes 2-7. Most of the high frequency charge sent to the current sources would be gotten from the lumped capacitor. The charge would have to encounter all of the impedance that lies after the capacitor in nodes 2-7. This is shown in Figure 9.

In Figure 9, there are 4 unique paths that can be examined to determine the maximum or average amount of look-ahead noise seen for the lumped strategy. The average amount of look-ahead noise provides an insight into how much noise could be

encountered if a specific decoupling strategy were followed. The maximum amount of look-ahead noise provides an insight into finding an extremely noisy path or node within the RLC tree. The equations for the lumped strategy in Figure 9 are:

$$\Delta V_{LA\_AVG1} = (\Delta V_2 + \Delta V_4 + \Delta V_2 + \Delta V_5 + \Delta V_3 + \Delta V_6 + \Delta V_3 + \Delta V_7)/4$$

$$\Delta V_{LA\_MAX1} = \max(\Delta V_2 + \Delta V_4, \Delta V_2 + \Delta V_5, \Delta V_3 + \Delta V_6, \Delta V_3 + \Delta V_7)$$

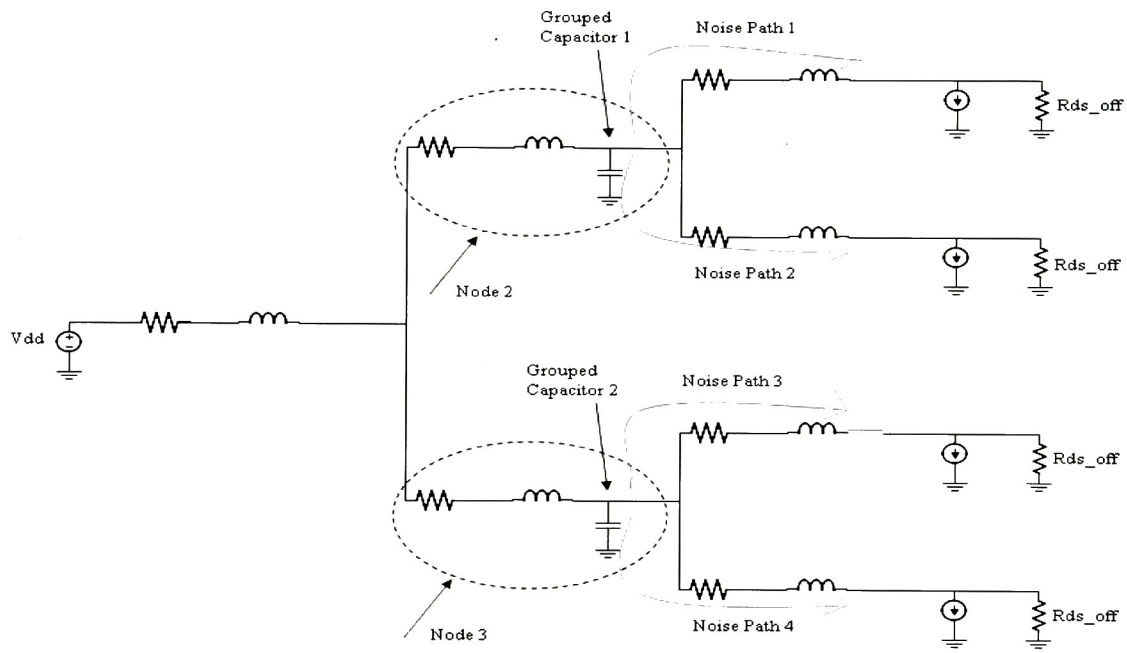

Figure 3.10: Grouped Capacitor Noise Paths

This can also be done for the grouped placement strategy. Again there are 4 unique paths to be examined but because the decoupling capacitance is placed on level 2, only nodes 4-7 factor in the noise fluctuations seen before the current sources.

$$\Delta V_{LA\_AVG2} = (\Delta V_4 + \Delta V_5 + \Delta V_6 + \Delta V_7)/4$$

$$\Delta V_{LA\_MAX2} = \max(\Delta V_4, \Delta V_5, \Delta V_6, \Delta V_7)$$

More generally, a variable j is defined to iterate over each distribution level. The variable i is defined to iterate over all unique paths. The vector RN holds all noise values for a specific placement strategy relative to all of its unique paths.

$$\begin{aligned} &\#\ unique\ paths = 2^{n-1} \\ &j = 2 \ldots n \\ &i = 1 \ldots 2^{n-1} \\ &RN_{j-1} = \sum_{k=j-1}^{n-1} N(i,k) \end{aligned} \tag{3.25}$$

$$\Delta V_{LA\_AVG\_j} = \left(\sum_{m=1}^{2^{n-1}} RN_m \right) \Big/ 2^{n-1} \tag{3.26}$$

$$\Delta V_{LA\_MAX\_j} = \max\ (RN) \tag{3.27}$$

3.4. Matlab Implementation

A *Matlab* script was written that implemented the approximation presented in this chapter. This was done to show how parameters could be extracted that would allow for interconnect comparisons of the different placement strategies. The procedure derived from Equations (6) to (20) was used to determine the amount of decoupling capacitance to use for different placement strategies. The metrics defined in Equations (22)-(23) and (26)-(27) were used as a guide to show possible differences between strategies.

The *Matlab* script took two input files as parameters. The first file characterized the distribution network as a symmetric binary RLC tree with negligible parasitic decoupling capacitance. The second contained the current data used to represent the switching activity of the digital blocks. The current sources were placed on the last level of the tree, one per node. These files were generated randomly with ability to be scaled to targeted values. For example, the RLC tree file could contain maximum interconnect

inductances of 25 pH, maximum interconnect resistances of 50 $m\Omega$, and maximum interconnect capacitances of 50 aF. The current model presented in 3.2.1.1 was used for the current sources within the RLC tree. For example, maximum values for the current data could be $30\,\mu A$ with a peak time of 50 ps. The script allows for an exact bondwire value to be entered at node 1 on the RLC tree.

Each of these maximum values is an input parameter and is randomly scaled using a random number that ranges from 0 to 1. For example, if in Figure 1 the targeted maximum resistance was $50\,m\Omega$, each of the 7 resistors within the RLC tree would be scaled by a random seed.

$$R_{1-7} = 50m\Omega * rand();$$

All of the preceding values are presented as examples to illustrate how the *Matlab* script functioned, they were not extracted from a specific technology.

Two different simulations were run on a 10 level tree with a 1 V supply. The first simulation had a λ of 5%, which correlates to allowed fluctuations of 50 mV or less and a minimum line voltage, $V_{dd\,\min}$, of 0.95 V. The second simulation had a λ of 10%, which correlates to allowed fluctuations of 100 mV or less and a minimum line voltage, $V_{dd\,\min}$, of 0.90 V The parasitic values used for the RLC trees in these simulations were extracted using TSMC 90 nm technology.

ASITIC, a CAD tool that models inductors, transformers, capacitors, and substrate coupling used a TSMC 90 nm model file to create the resistance and inductance values [29]. For example, Figure 9 shows an example of how *ASITIC* is used. Once the model file has been loaded, the user can create layout type structures (spiral, square, ring, wire,

etc) in a grid based display. The user can then run commands from a terminal on these structures to determine associated inductances, resistances, and capacitances.

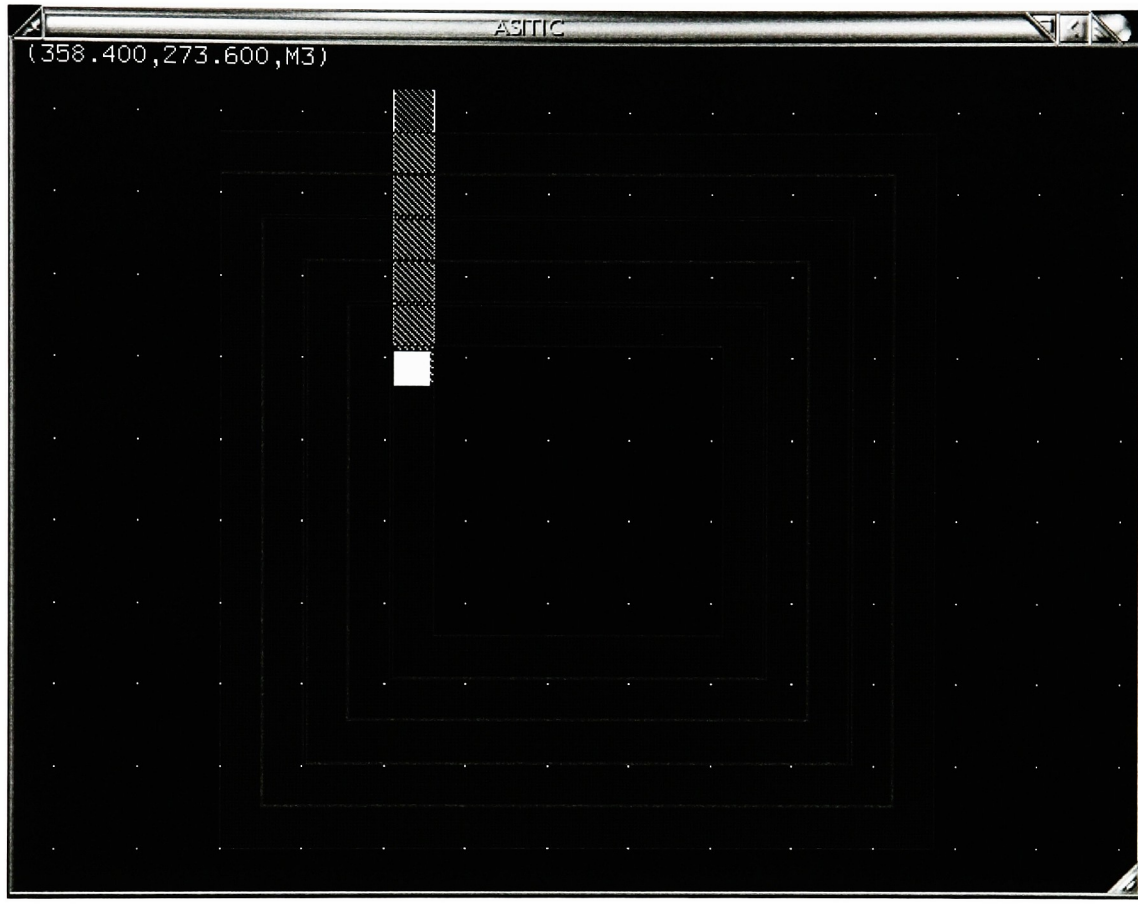

Figure 3.11: *ASITIC* Display

A small digital circuit using the TSMC 90 nm technology was simulated in *Cadence* to model the current sources that were placed on the last level of the tree. The schematic and logic function are shown in Figures 12 and 13.

$$C = \left((X \bullet Y)' \oplus Mode\right)'$$

$$Sum = A \oplus (B \oplus C)$$

$$Carry\_Out = B \bullet C + A \bullet C + A \bullet B$$

Figure 3.12: Digital Test Circuit Logic Function

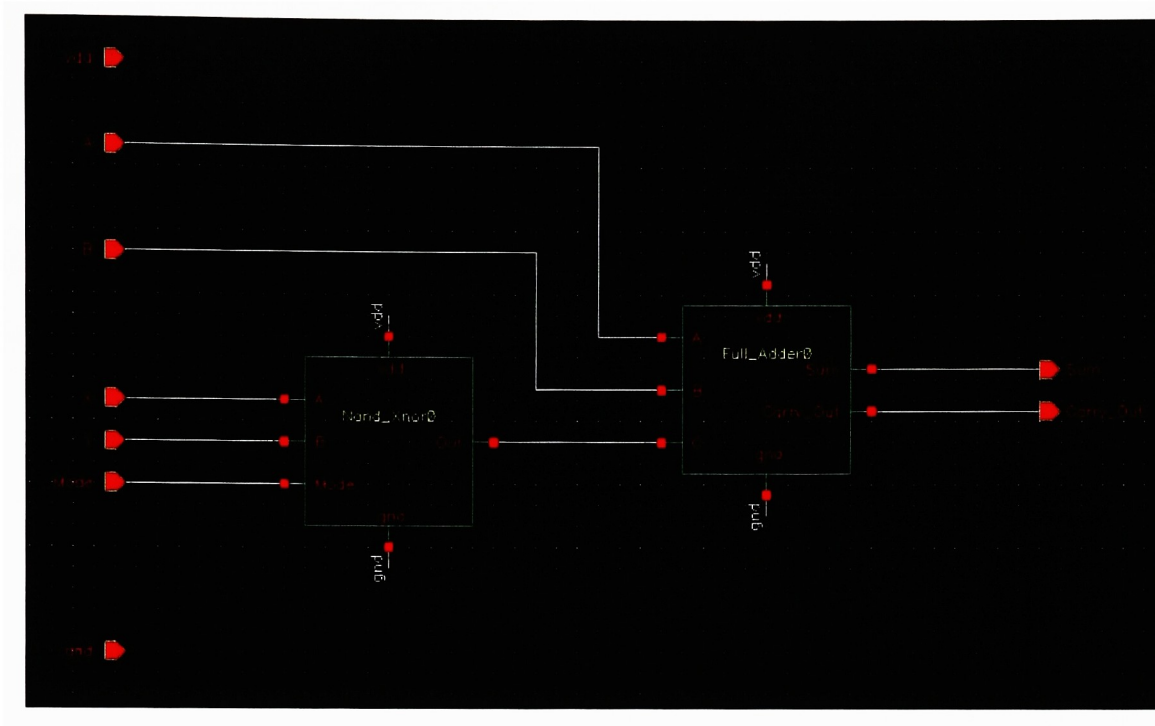

Figure 3.13: Digital Test Circuit Schematic

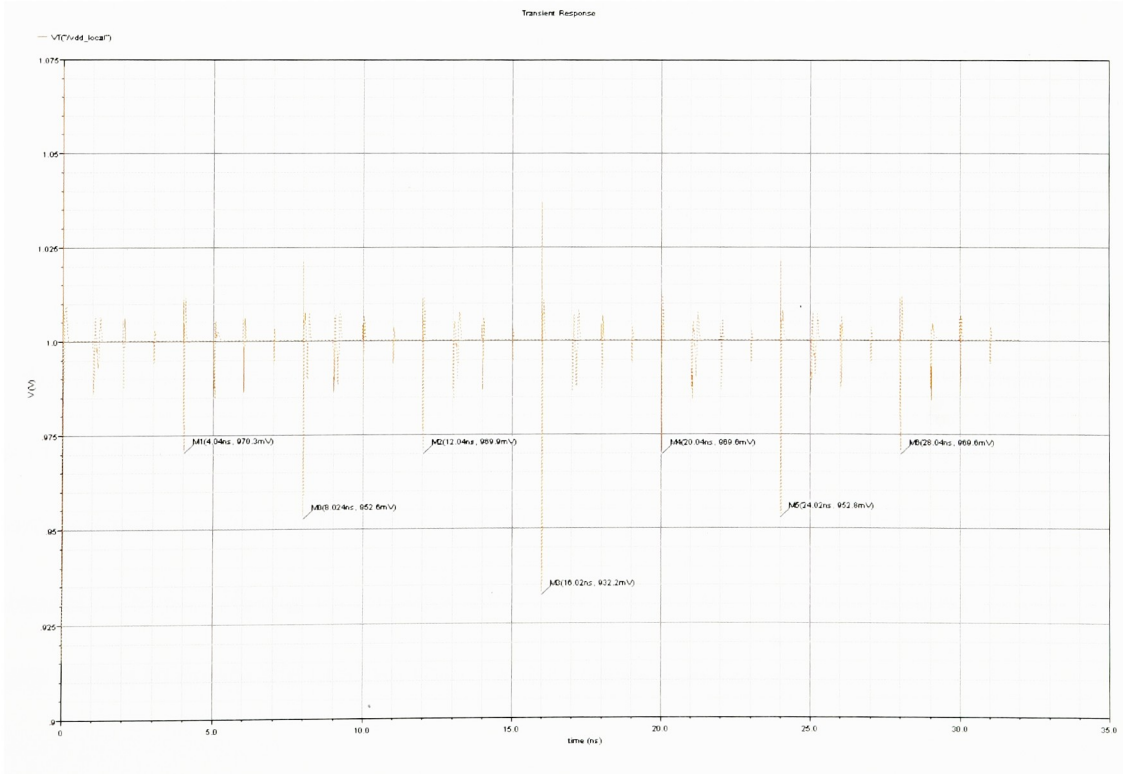

Figure 3.14: Digital Test Local Supply Line

A simulation was ran where each input was switched every 2 ns, which correlates to a 500 MHz clock. A single 5 nH bondwire was placed between the 1 V power supply and the logic. The length of simulation was 32 ns, enough time to allow all possible input combinations. Figures 14 and 15 show the local power supply for the logic block after the bondwire and the amount of current drawn through the bondwire.

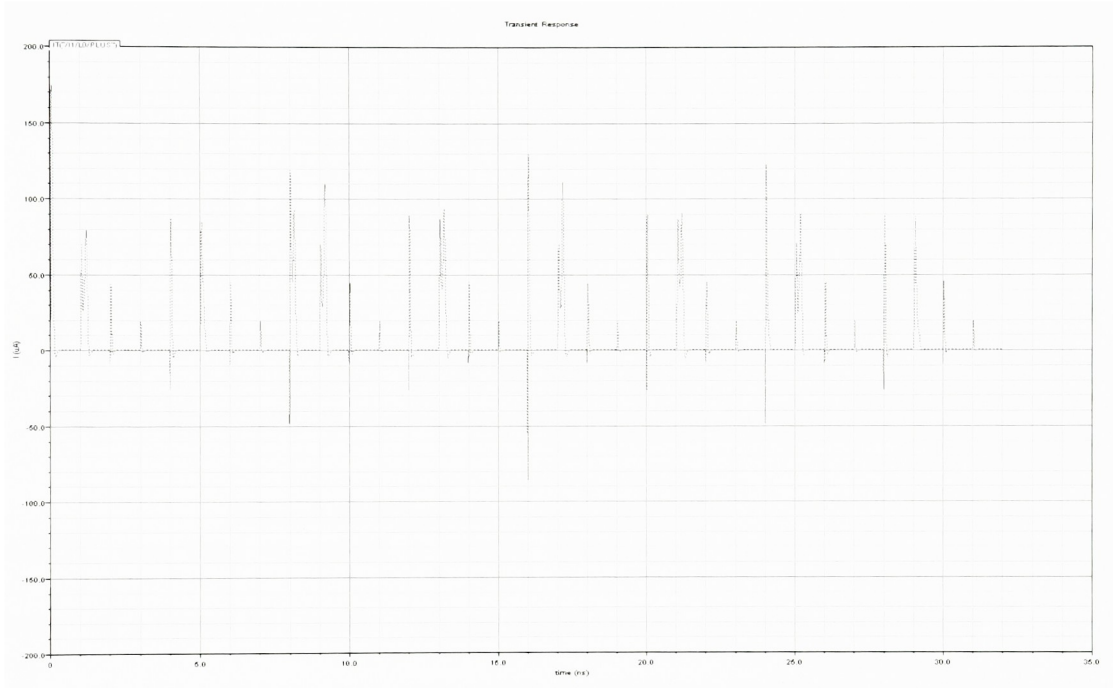

Figure 3.15 Digital Test Current Draw

In order to extract the two parameters needed to create the current parameter file for the *Matlab* script, the plot in Figure 16 was used. This plot shows the maximum amount of current over the longest period of time needed during the simulation. The maximum current is approximately 125 μA and the switching activity lasts for approximately .26 ns. These two values directly correlate to what was entered as the maximum values to create the random data for the current sources.

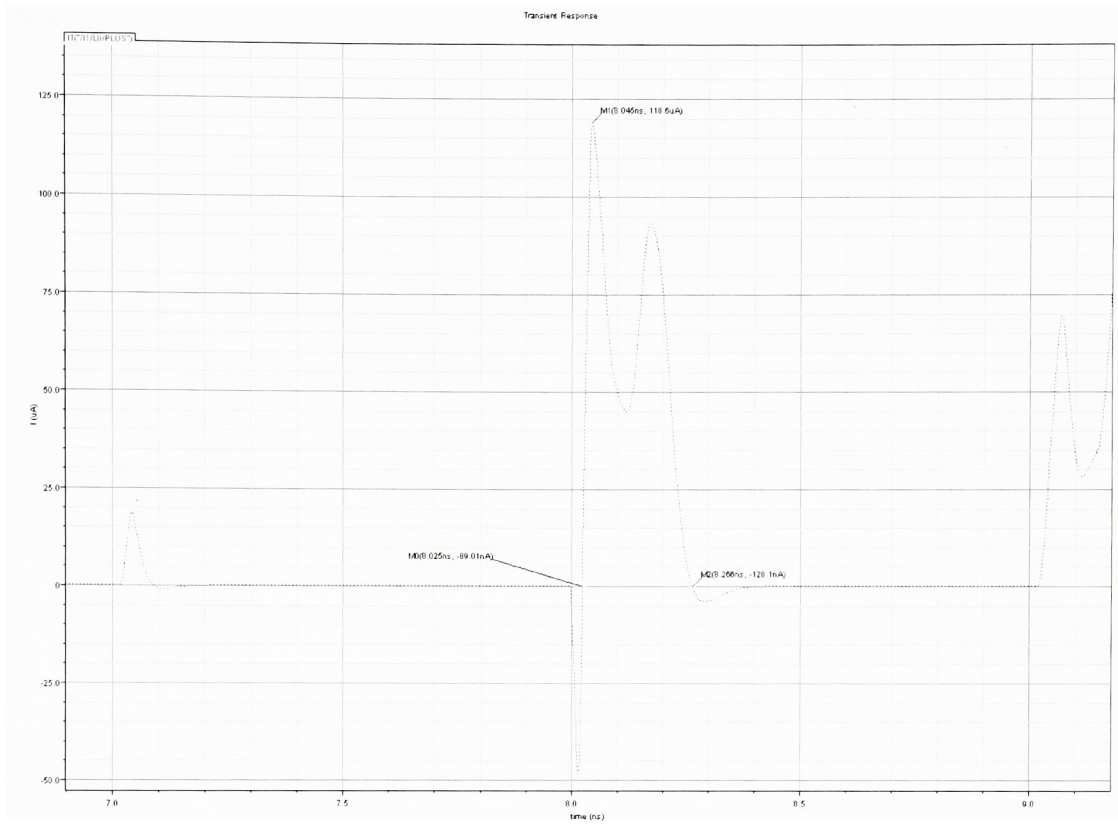

Figure 3.16: Magnified Digital Current Plot

ASITIC was used to model a simple interconnect structure that might be representative of a power line within a 90 nm digital circuit. Because this structure was going to be randomly scaled on every node within the RLC tree, its length and width was kept minimal in order to avoid creating unrealistic inductance and resistance values. A wire structure joining metals 1-8 with a length of 50 μm and a width of 2 μm modeled the resistance and inductance within the RLC tree. Shown in Figure 17, the analysis gave an inductance of 50.38 pH and a resistance of 500.3 $m\Omega$. These two numbers directly correlated to the values that were used as maximums in order to create the RLC input file mentioned above. The capacitance within this structure was negligible and a maximum value of 0.1 fF was used to create the RLC input file.

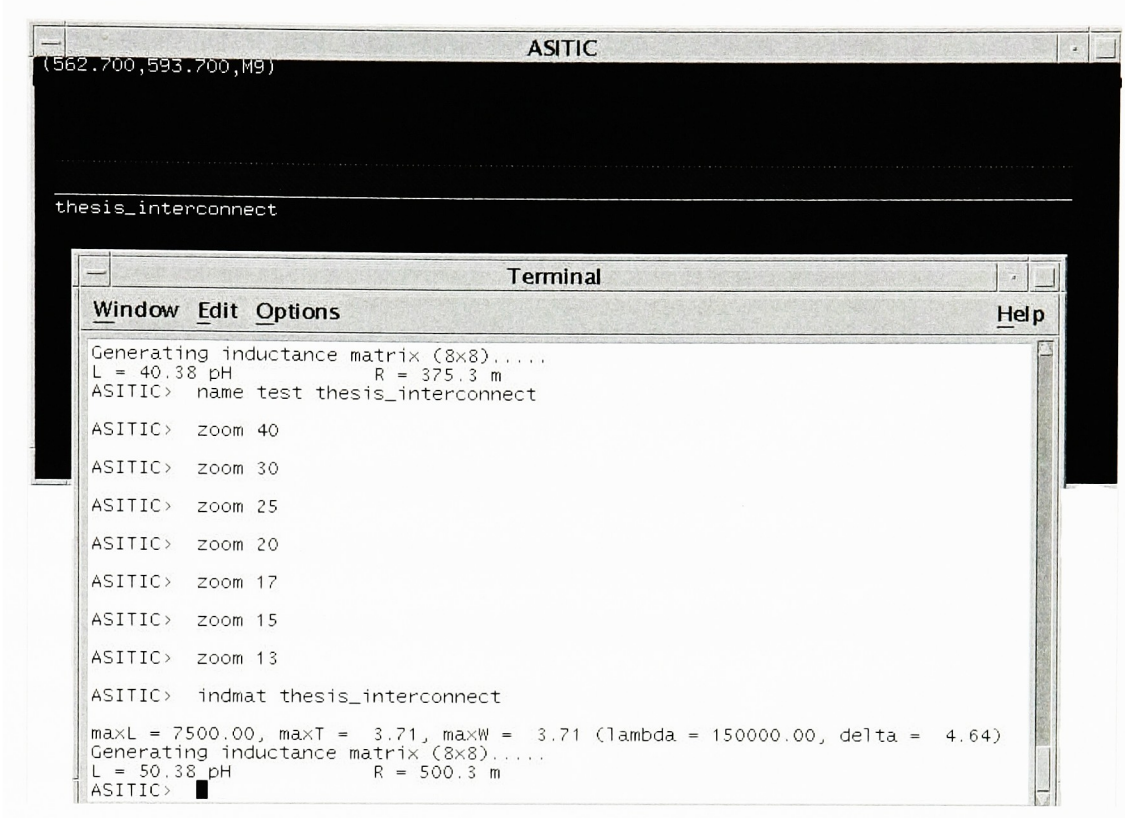

Figure 3.17: *ASITIC* Interconnect Model

Table 1 shows the range of values for the two input files. Resistance, inductance, and capacitance values within the ranges depicted in the table were placed on each node within the 10 level tree. The current sources on the last node of the tree had current peak parameters ranging from 0.4 μA to 124.9 μA and peak times ranging from 0.24 ps to 129.75 ps.

Component	Mean	Min	Max	Std. Dev.
Resistance ($m\Omega$)	248.967120	0.477426	499.695545	146.346007
Inductance (pH)	25.566348	0.002525	49.869335	14.360681
Capacitance (fF)	0.051680	0.000152	0.099863	0.028737
Current Peak (mA)	0.061734	0.000485	0.124910	0.037164
Peak Current Time (ps)	60.744363	0.241589	129.751270	38.437784

Table 3.1: Interconnect Parameter Ranges

The *Matlab* script also implemented the theory presented in [9]-[11]. This was done to extract the second order specifications mentioned in the Chapter 2 and presented in [9]-[11]. Each decoupling strategy has different specifications, which can be compared. For example, if the natural frequency of one decoupling strategy is much greater than another decoupling strategy, it may imply one strategy has advantages over the other.

The following three sub-chapters show the simulation results for the different activity factors. For each table or figure, the term distribution level implies the level of decoupling capacitance distribution. Distribution level going from 1 to 10 is analogous to going from completely lumped to completely distributed. The first sub-chapter summarizes the interconnect comparison results. The second sub-chapter presents the noise metrics. The third sub-chapter attempts to identify the optimal level of distribution for each simulation.

3.4.1 Interconnect Comparison

The results in this sub-chapter are presented in following manner. For each pair of figures shown, the first figure is from the simulation with an activity factor of 5% and the second figure is from the simulation with an activity factor of 10%.

The analysis is done on a single node within the RLC tree, as per the work presented in [9]-[16]. The *Matlab* script accepted the target node as a parameter. For example, in the 10 level tree, any of the 1024 nodes could be targeted. If node 796 was chosen, second order parameters would be extracted for that specific node. It is important to note that the script calculates the amount of capacitance needed at each node for a single distribution level. The analysis is then done on the targeted node for that

distribution level to extract the second order parameters ands specifications. This is done for all levels of distribution, from lumped to distributed. For example, if the script was doing the analysis on the 3rd level of distribution, it would first calculate capacitance values for each node within the third level. It would then perform the analysis presented in [9]-[11] on the targeted node. The results of that targeted node would change when the level of distribution changes. Pseudo code is shown in Figure 18.

```
//this function performs the second order analysis for each level in the
 //tree on a targeted node
allResults = doSecondOrderAnalysis(numLevels,targetNode,RLCtree){

//go through each level of distribution
for(i = 0; i < numLevels; i++){

//get all of the nodes in this level
thisLevelNodes = getLevelNodes(i+1,RLCtree);

//put the decoupling capacitance into these nodes
putDeCap(thisLevelNodes,RLCtree);

//perform second order analysis for this level on the target node
[singleLevelResults] = secondOrderAnalysis(RLCtree,targetNode);

//store this levels results
allResults[i] = singleLevelResults;

//clear capacitance from nodes in this level
clearDeCap(thisLevelNodes,RLCtree);
}
}
```

Figure 3.18: Second Order Analysis Pseudo-Code

Node 511 was chosen for this analysis. Many simulations were run on different nodes and the second order results were comparable.

Tables 2 and 3 show the second order parameters extracted from the analysis. In both simulations, settling time and overshoot improve as distribution level increases.

This leads to the notion that the supply would be more stable in a higher level distribution. Peak time, rise time, and delay time all become worse as distribution level increases. Delay time is an important specification because it models the amount of time a signal would need to propagate an interconnect line. As delay times increase, the ability to clock circuitry at high frequencies reduces. This result warrants the use of a distribution strategy closer to lumped.

Distribution Level	Peak Time (ns)	%Overshoot	Settling Time (ns)	Rise Time (ps)	Delay Time (ns)
10	2.46	96.91	235.08	837.05	0.84
9	2.42	96.95	235.08	824.02	0.82
8	2.36	97.05	236.18	799.79	0.80
7	2.32	97.12	238.82	786.91	0.79
6	2.24	97.26	242.25	757.00	0.76
5	1.93	97.87	270.03	647.05	0.65
4	1.81	98.15	291.57	603.87	0.61
3	1.79	98.31	315.33	594.59	0.60
2	1.72	98.82	434.03	568.58	0.58
1	1.41	99.65	1915.50	458.19	0.47

Table 3.2: 5% Activity Factor Specifications

Distribution Level	Peak Time (ns)	%Overshoot	Settling Time (ns)	Rise Time (ps)	Delay Time (ns)
10	1.19	98.48	235.07	395.52	0.40
9	1.19	98.49	235.07	394.24	0.40
8	1.18	98.50	236.17	391.60	0.39
7	1.17	98.53	238.80	390.09	0.39
6	1.16	98.56	242.23	386.34	0.39
5	1.12	98.76	269.95	369.29	0.37
4	1.09	98.87	291.45	360.83	0.36
3	1.09	98.96	315.15	358.67	0.36
2	1.07	99.25	433.47	352.24	0.36
1	0.98	99.75	1184.50	321.26	0.33

Table 3.3: 10% Activity Factor Specifications

Figures 19 and 20 show the amount of capacitances needed to meet activity factor requirements for each distribution level. It should be noted that these capacitance values represent the total amount inserted into the entire tree, not just the targeted node. The

amount of capacitance needed per level of distribution increases as placement moved from level 1 to level 10 in both figures. This was due to the resistive losses in the interconnect itself. The increase in capacitance in Figure 19 (5% activity factor) was far greater than the increase in capacitance in Figure 20 (10% activity factor). It can be concluded from this result that stringent activity factors correlate to high values of decoupling capacitance which lead to an increased amount of on-chip area.

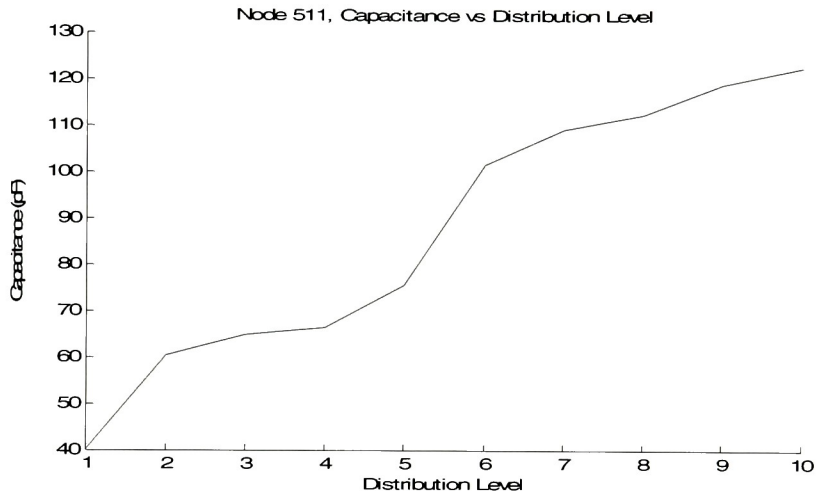

Figure 3.19: 5% Act. Factor Distribution Level vs. Total Capacitance

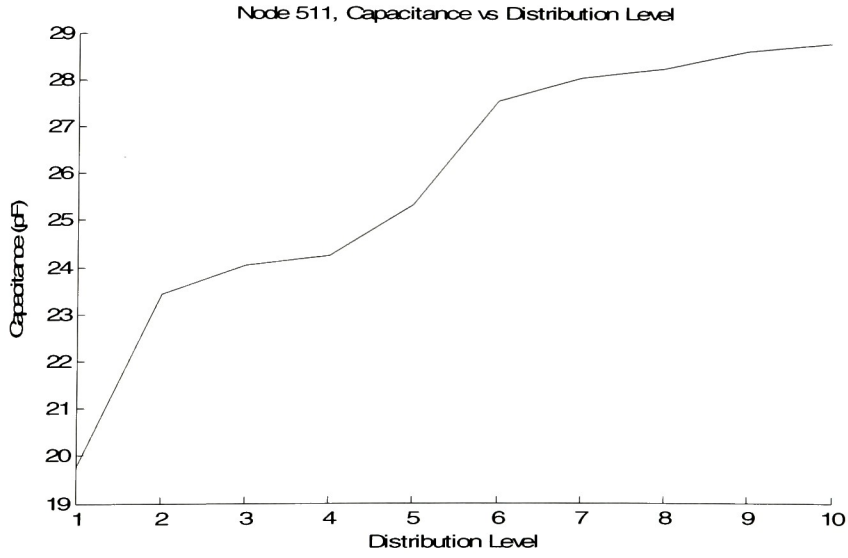

Figure 3.20: 10% Act. Factor Distribution Level vs. Total Capacitance

Figures 21 and 22 show the damping factor results for each activity ratio simulation. In both figures, the damping factor increases as distribution level increases. The damping factor for the 5% activity factor simulation are approximately doubled those of the 10% activity factor simulation. This implies that more capacitance (5% activity factor) is related to a larger damping factor. In a second order system, the damping factor is defined as:

$$\xi = \frac{R}{2L}$$

There is no relationship to capacitance in the above formula. In the work presented in [9]-[11] it was shown the damping factor has a dependence on capacitance; increasing capacitance leads to a larger damping factor. As damping factors increase, oscillations decrease.

It should be noted that in each simulation, all levels of distribution were underdamped. This was expected because of the simple second order relationship for underdamping, shown in the following equation. From this equation it can be seen why the interconnect structures were underdamped for each level of distribution; the overall amount of capacitance was far greater than the amount of inductance.

$$RC < \frac{4L}{R}$$

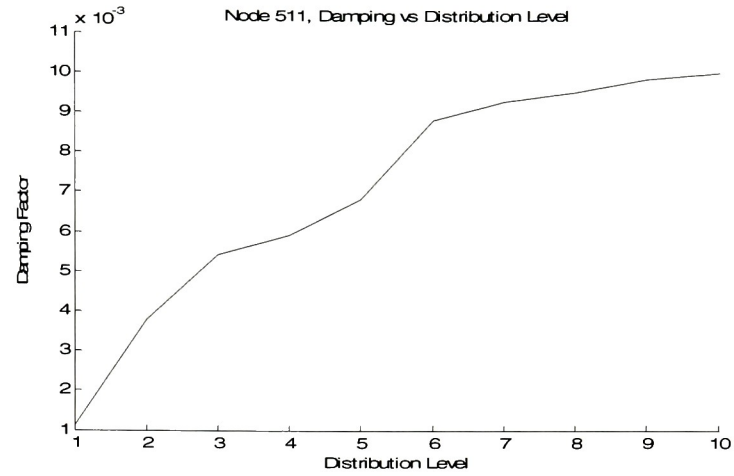

Figure 3.21: 5% Activity Factor Distribution Level vs. Damping

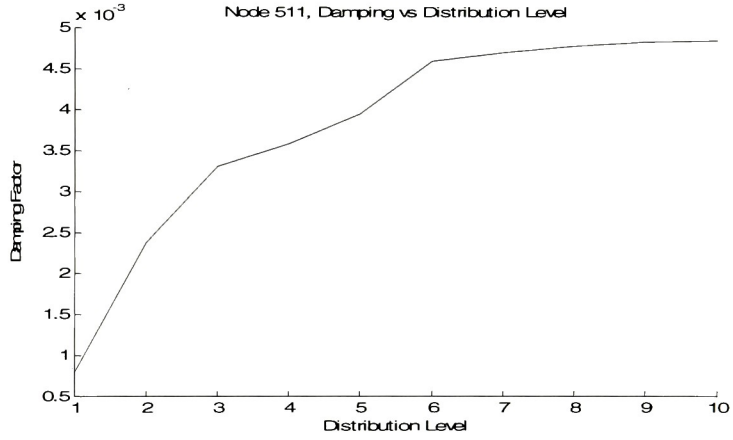

Figure 3.22: 10% Activity Factor Distribution Level vs. Damping

Figures 23 and 24 show the natural frequency results for each simulation. In both cases, natural frequency decreased as distribution increased. For a second order system, the natural frequency is defined as:

$$\omega_n = \frac{1}{2\pi\sqrt{LC}}$$

The increase in natural frequency in both Figures was expected due to the increase in capacitance per level. The natural frequencies in the 5% activity factor simulation are approximately half of what is shown in the 10% activity factor simulation.

This result shows that as the activity factor becomes smaller, the amount of capacitance increases causing the natural frequency to decrease.

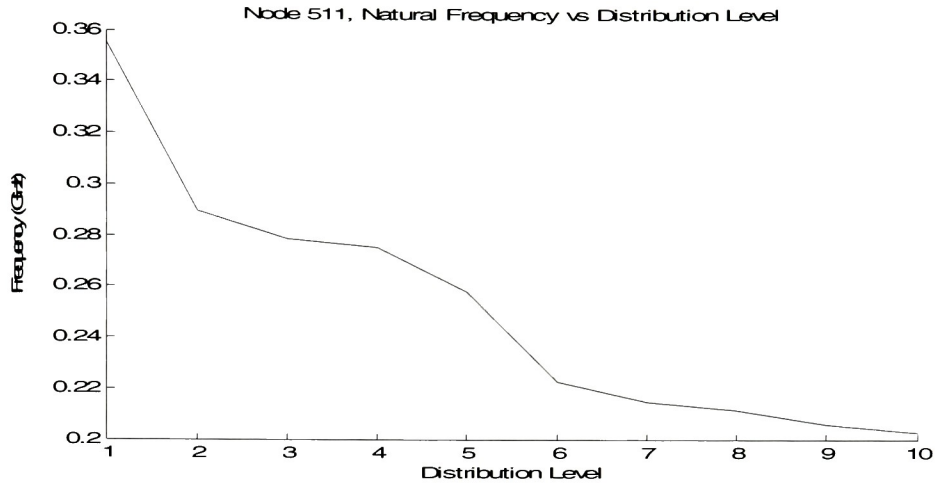

Figure 3.23: 5% Activity Factor Distribution Level vs. Natural Frequency

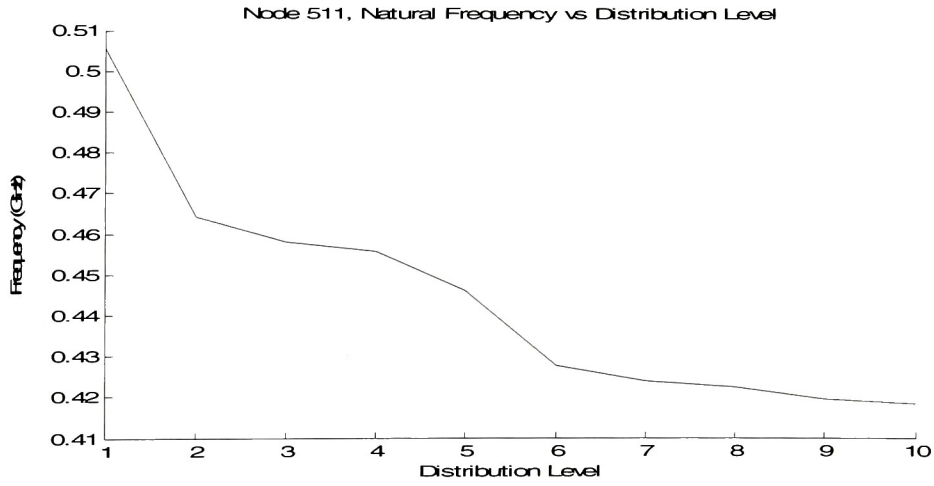

Figure 3.24: 10% Activity Factor Distribution Level vs. Natural Frequency

The simulations presented in this sub-chapter show the trade-offs between different levels of distribution and second order specifications relative to an activity factor. If a design has a minimum amount of on-chip area, it could target a higher activity factor or a lower level of distribution in order to reduce the amount of decoupling

capacitance. If rejecting supply noise is more important, a lower activity factor or higher level of distribution may be more applicable.

3.4.2 Noise Metrics

The follow figures show the noise metric calculations described in sub-chapter 3.3. The same RLC tree was used for both simulations, which explains why there are only one set of plots.

Figures 25 and 26 show the average and worst case amount of inductive level noise. It is important to notice that distribution level 1 is not shown on either figure. This is because level 1 represents the bondwire, which without any decoupling capacitance would produce an extremely high amount of noise. If this value was included in either figure, it would dominate the other values and make them incredibly difficult to see. The simulation results showed that the bondwire is capable of producing 9.23 V of noise. This is the average and worst case value because the bondwire is the only node in level 1 of the RLC tree. In Figure 25, the general trend is the average amount of noise per level decreases as distribution increases. This is not true in distribution level 3, where the average level noise rises from approximately 6 mV to 13.5 mV. This could imply a noise hot spot within that level. This is confirmed in Figure 27, where level 3 has a worst case noise value of 24 mV. Also, in Figure 26, the worst case noise results do not follow the same trend seen in Figure 23. Levels 8 and 10 have larger worst case noise fluctuations that some of their preceding levels. Although these values are not extremely high (~5 mV), it proves that there may be certain nodes within the RLC tree capable of producing an excessive amount of noise. Excessive noise

can be attributed to a large interconnect inductance at a single node, a large amount of required current, or high switching rates.

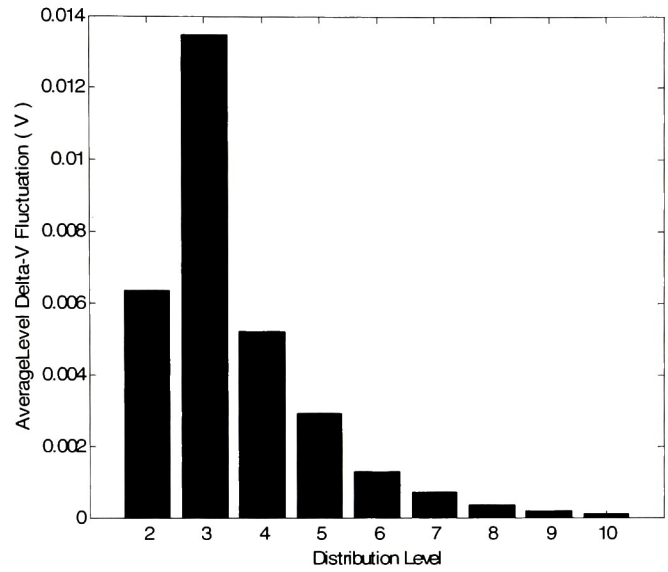

Figure 3.25: Distribution Level vs. Average Inductive Noise

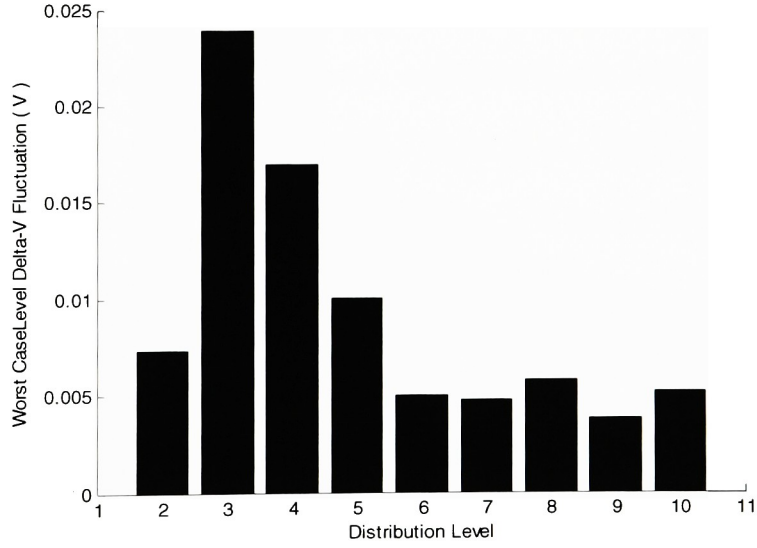

Figure 3.26: Distribution Level vs. Worst Case Inductive Noise

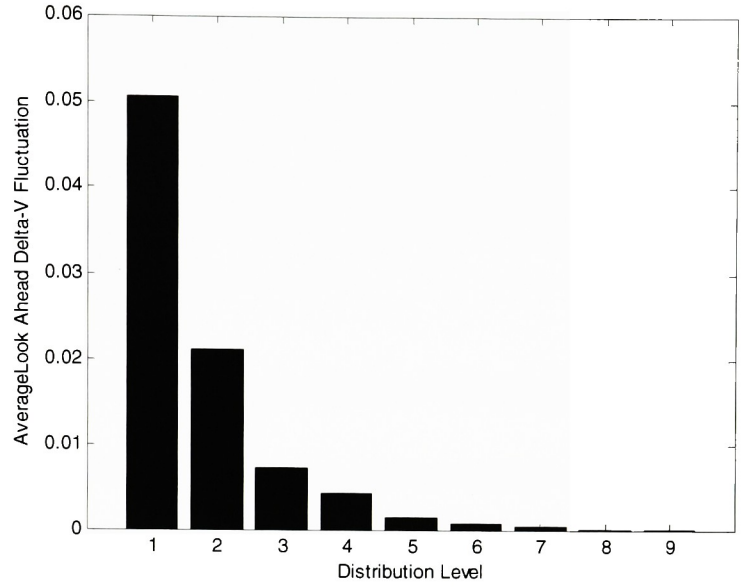

Figure 3.27: Distribution Level vs. Average Look-Ahead Inductive Noise

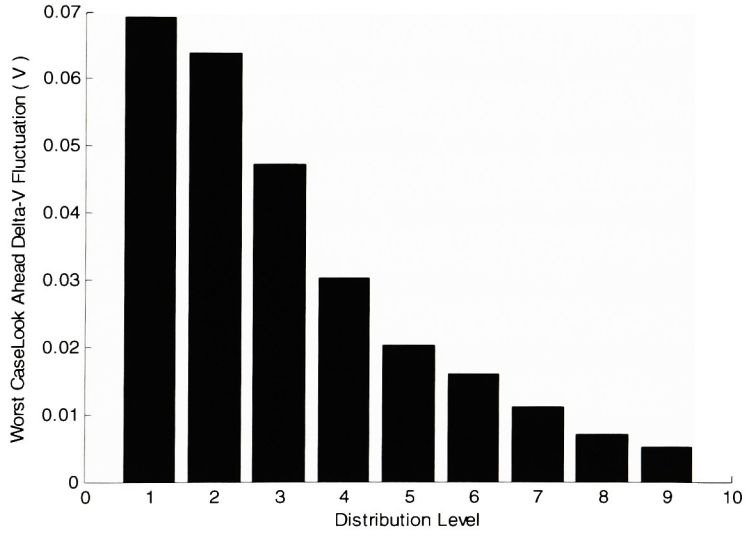

Figure 3.28: Distribution Level vs. Worst Case Look-Ahead Noise

Figures 27 and 28 show the average and maximum amount of look-ahead noise. Both figures show trends of decreasing look-ahead noise as distribution level decreases. This was expected due to how the calculation is performed; noise values are summed starting from a beginning node to an end node. In Figure 27, if a level 1 (lumped) strategy was followed the graph shows that there is an average of 50 mV along each

forward path after the lumped capacitor. In Figure 28, the worst case look-ahead noise path for a level 1 strategy is 70 mV. Both graphs show that the solution to decreasing look-ahead noise is to distribute capacitance closer to the switching elements. Distribution level 10 is not shown on either plot because there are no inductors after any node in the last level of the RLC tree.

3.4.3 Optimal Decoupling Capacitance Placement

In order to determine where to place the decoupling capacitance, different factors must be taken into account.

First, the placement strategy must be able to meet activity factor requirements. This can be done by looking at the noise metric figures in the previous section. For the activity factor of 10%, which allows 100 mV of fluctuation, any level of distribution is acceptable. This is not the case for the simulation with an activity factor of 5%, which allows only 50 mV of acceptable supply noise. In Figure 28, the average amount of look-ahead noise is approximately 50 mV for a level 1 distribution. This implies that at least a level 2 distribution would need to be followed. In terms of worst case situations, Figure 29 shows a relatively high amount of possible look-ahead noise on levels 1-3. If this particular design was extremely concerned with supply noise, at least a level 3 distribution would be followed.

Second, once the placement strategy has been selected the second order parameters for that specific level of distribution can be examined. If a lumped strategy was selected for the 10% activity factor simulation, the amount of capacitance needed would be approximately 20 pF. The natural frequency and delay time would both be optimal at 500 MHz and 0.33 ns. If a level 4 distribution strategy was selected for the 5%

activity factor simulation, the amount of capacitance needed is approximately 65 pF. This requires an extra 25 pF compared to a lumped strategy. The natural frequency is also quite low at approximately 270 MHz. The delay time is 0.61 ns. These results show that if tight control of noise is targeted, many different trade-offs occur; amount of capacitance needed will increase, delay time will increase, and natural frequency will decrease.

Chapter 4 Testing Strategy

A complete digital design was done using TSMC 90 nm technology to test some of the concepts presented in Chapter 3. The requirements of the design were that it contained a practical digital structure capable of constantly switching. A high amount of switching ensures a large high frequency current draw capable of producing excess supply noise. Without a reasonable amount of supply noise, differences between decoupling strategies cannot be shown.

Figure 1 shows a voltage transfer curve of an inverter for the TSMC 90 nm process. The high noise margin, V_{nmh}, was found to be 470 mV. The low noise margin, V_{nml}, was found to be 430 mV. Both of these noise margins are extremely high considering the supply is only 1 V.

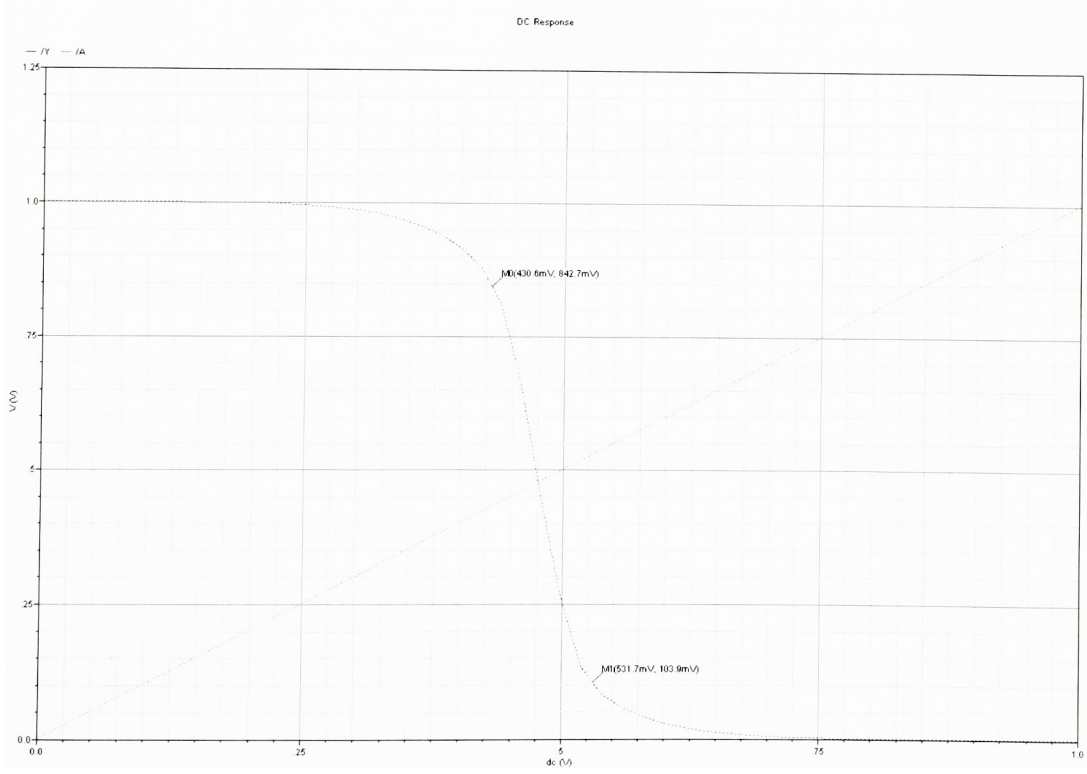

Figure 4.1: Inverter Voltage Transfer Curve

This chapter is broken down into 4 sub-chapters. The first sub-chapter explains the high level functionality of the noise generation circuitry. The second sub-chapter presents layout techniques used to implement the design. The third sub-chapter presents simulation results used to size the decoupling capacitors. The fourth sub-chapter discusses what needs to be done to monitor the on-chip noise.

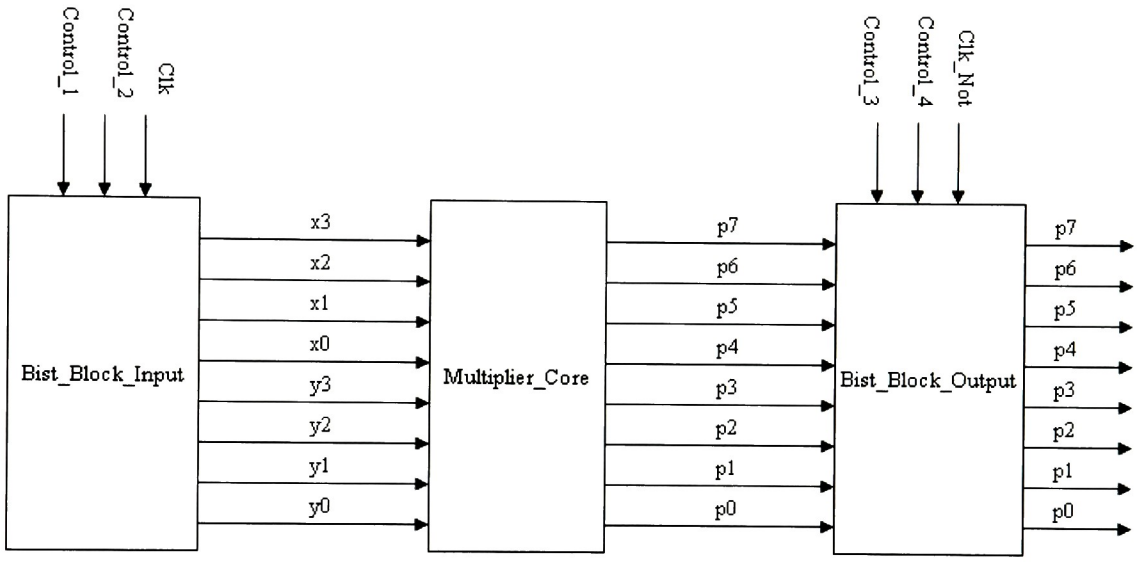

Figure 4.2: Multiplier Block

4.1. Noise Generation Circuitry

A digital 4x4 multiplier was chosen as the basic digital block because it met all of the above requirements. The basic multiplier structure consisting of a multiplier core and two BIST blocks is shown in Figure 2.

The entire block is controlled by five input signals; four DC control lines and one clock. The control lines for the input BIST block are responsible for generating pseudo-random input bits. These bits are fed to the multiplier core, which performs a signed 4x4 multiplication. The output bits from the multiplier core are then fed into the output BIST

block. The output BIST block creates a unique signature out of each set of product bits. Control lines 3 and 4 are responsible for placing the output BIST in signature generation mode.

Depending on the inputs to the multiplier core, the switching state of the individual digital blocks within it will be different. Also, each BIST block is constantly changing states due to the random pattern generation. As long as the control lines are set in the appropriate manner, a continuous cycle of pseudo-pattern generation, multiplication, and signature analysis will occur. This creates a continuous pseudo-random high frequency current draw. This is not only ideal for supply noise generation, but also representative of a real digital circuit that is constantly stressing the power supply by changing states.

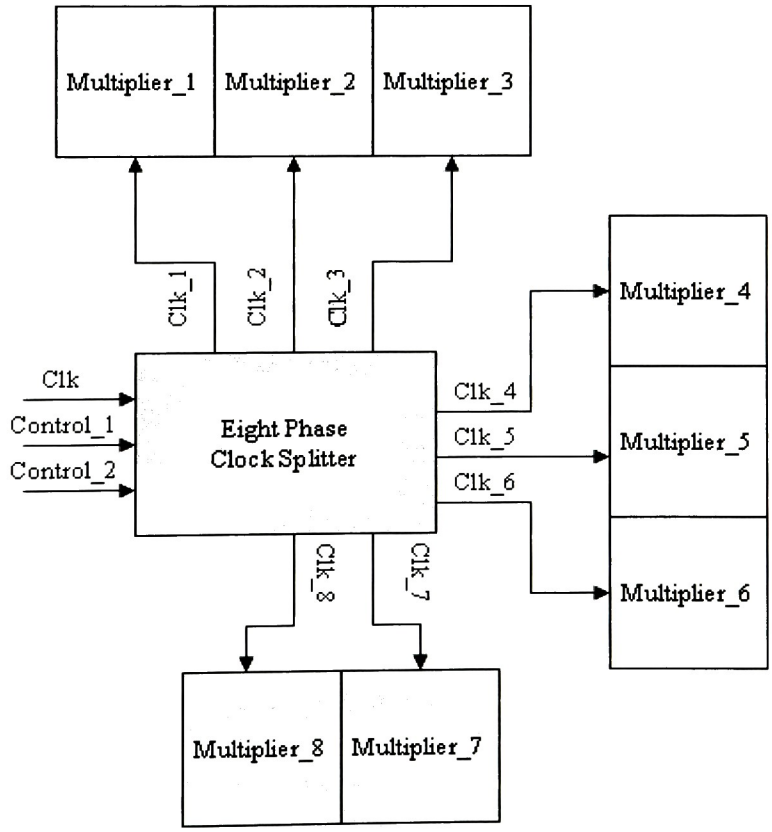

Figure 4.3: Multiplier Module

A single multiplier block is not capable of producing enough supply noise to test different decoupling strategies. To generate more noise, eight multiplier blocks were used to create a multiplier module, shown in Figure 3. The DC control lines are omitted from each multiplier for the purpose of visual clarity.

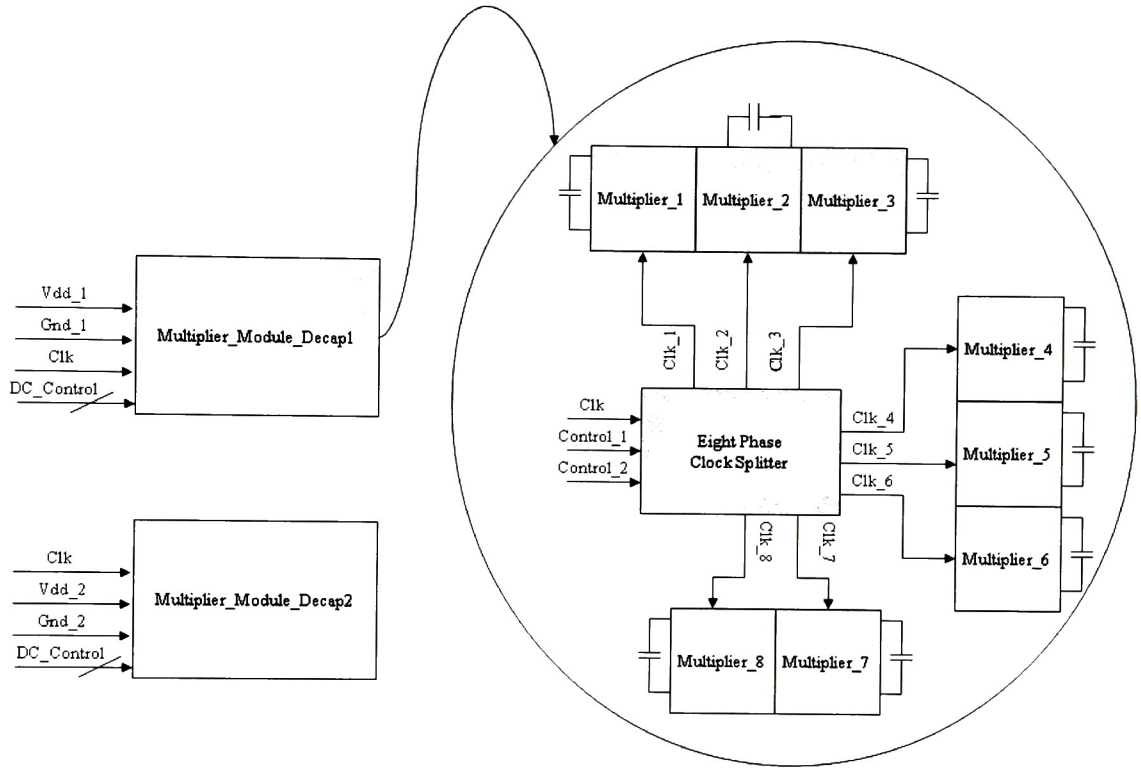

Figure 4.4: Separate Multiplier Modules

The module is driven by two DC control lines and a single clock. The block titled "Eight Phase Clock Splitter" is a specially designed digital block capable of dividing an input clock into an output clock with 8 different phases. If each multiplier module switches on the same clock edge, then a maximum amount of supply noise will be generated because a maximum amount of high frequency current will be flowing through the on-chip interconnect structure. On the other hand, if the input clock is divided into eight phases then all switching will not occur on the same clock edge thus limiting the

amount of supply noise. This technique allows some control over how much noise will be produced within a multiplier module. The two control lines are used to determine how the input clock is going to be divided. The settings are shown in Table 1. For example, when Control_1 and Control_2 are both low, each multiplier is fed a different phase of the input clock. When each control line is high, each multiplier is fed the same phase of the input clock. It is important to note that in order to get the desired phase division, frequency division also had to occur. The input clock frequency was divided by 4 in order to get the appropriate phases. For example, if the input clock was 2 GHz and both control lines were low, there would be 8 clocks with different phases, each with a frequency of 500 MHz.

Control_1	Control_2	Phase Division
0	0	8 different phases
0	1	4 different phases
1	0	2 different phases
1	1	1 phase

Table 4.1: Clock Phase Splitter Configuration

A different decoupling strategy can be implemented within a single multiplier module. For example, if two modules are placed on chip, one of them can implement a lumped strategy while the other implements a distributed strategy. This high level idea is depicted in Figure 4. Note that there are no physical connections between Multiplier_Module_Decap1 and Multiplier_Module_Decap2. This is purposely done to isolate different decoupling strategies. If in Figure 4 each module were connected to the same supply, capacitance from one module may help reject supply noise for the other module. If this occurred, it would make it impossible to test the effectiveness of the different decoupling strategies.

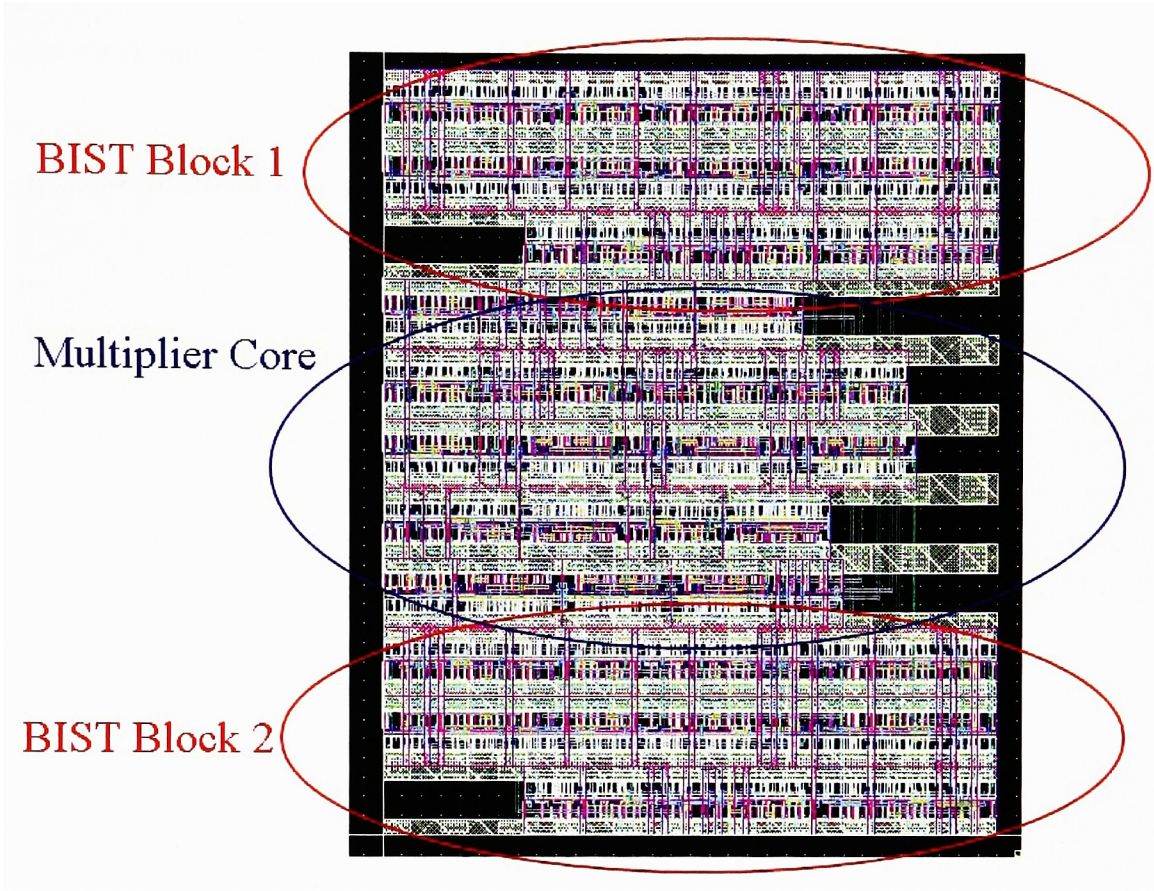

Figure 4.5: Multiplier Block Layout

4.2. Physical Layout

The physical layout of the noise generation circuitry was completely manual. The design started with implementing primitive gates. (NAND, NOR, XNOR, D Flip Flop, etc). Once these primitive gates were completed, they were used to build a multiplier block (Figure 5) and the eight phase clock splitter (Figure 6).

Eight multiplier blocks and the clock phase splitter were used to build a multiplier module (Figure 7). It is important to note that the clock phase splitter is in the middle of the multiplier module layout. This was purposely done to ensure similar delays between

each output clock pin of the phase splitter and its corresponding multiplier block input clock pin.

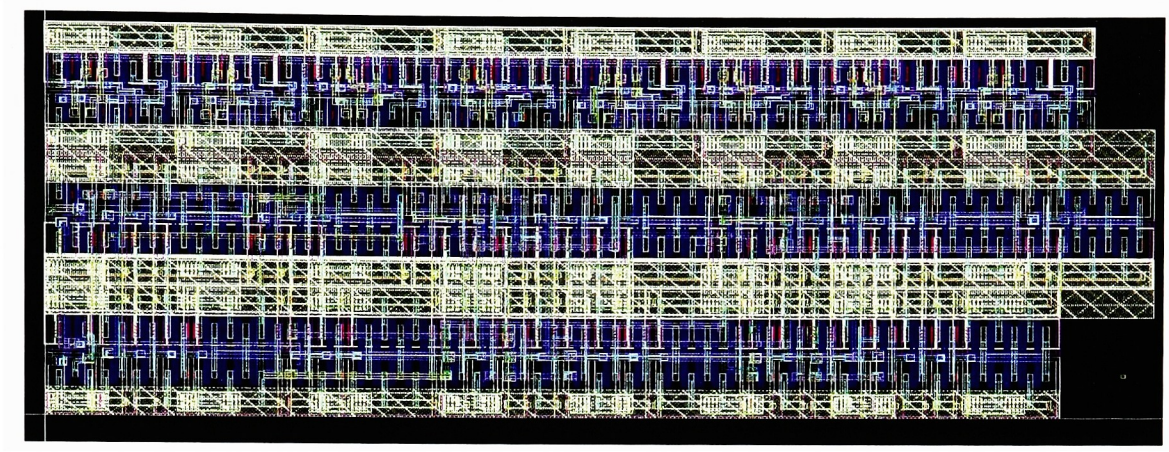

Figure 4.6: Eight Phase Clock Splitter Layout

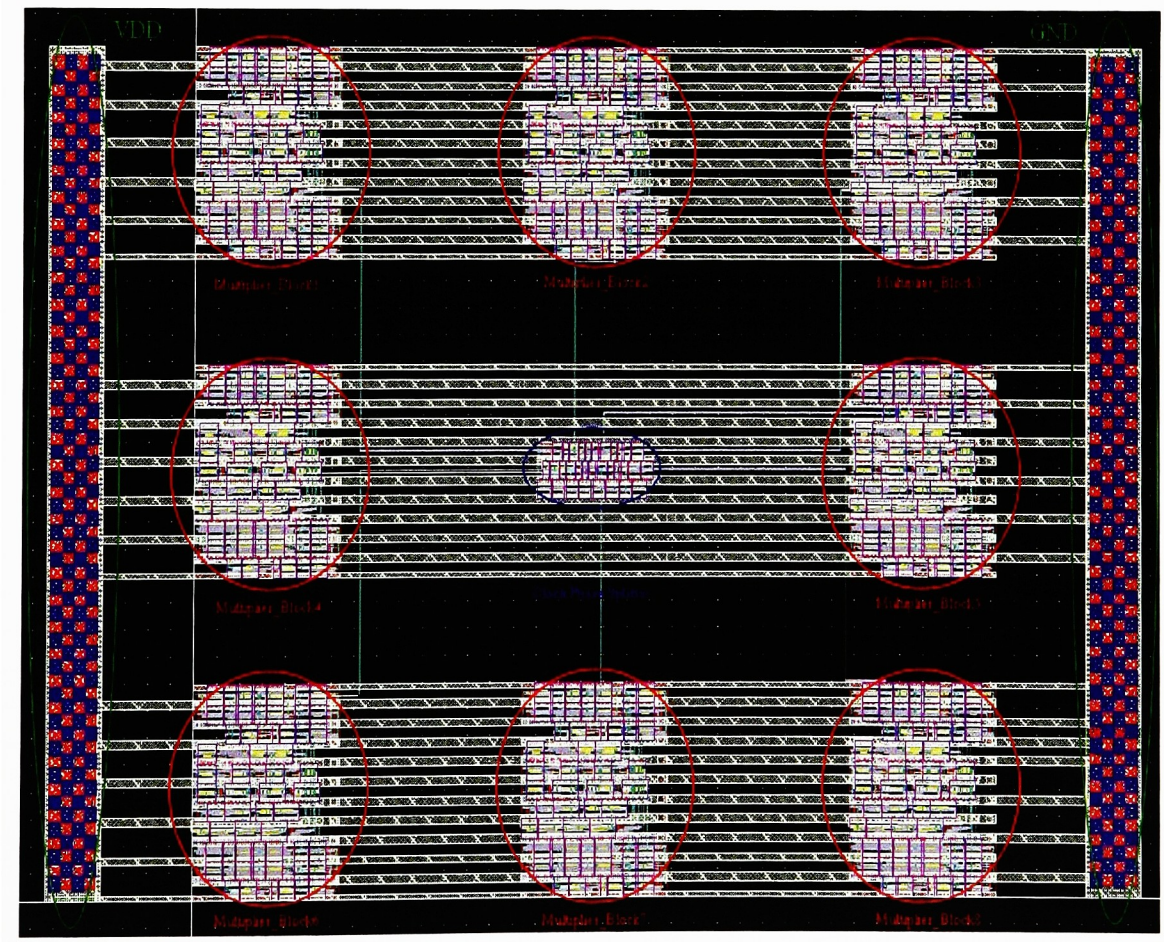

Figure 4.7: Multiplier Module Layout

Notice how the interconnect power bus is routed on the multiplier module. The thick vertical metal connection on the left hand side is the power input. The thick vertical metal connection on the right hand side is the ground input. The thin horizontal metal wires were used to correctly connect the individual multiplier blocks and clock phase splitter to power and ground. Metals 1-8 were layered wherever possible within the power interconnect to limit parasitic resistance. This was done to model how things are done in industry. The large empty spaces in multiplier blocks were left for the placement of decoupling capacitors.

4.3. Simulation Results

Once the layout of the multiplier module was complete, parasitic resistances and inductances were extracted using *ASITIC*. In Figure 7, there are four general types of interconnect wire structures. They have been highlighted in Figure 8. These four structures were used to model the interconnect of a single multiplier module. Table 2 shows the characteristics of each structure. Wire structure I1 was added to represent impedance between the last line of interconnect and terminals of the transistors.

Name	Inductance (pH)	Resistance ($m\Omega$)	Width (μm)	Length (μm)	Metals
V1	19.3	47.62	12	40	2-8
V2	7.428	56.22	12	40	3,5,7
H1	22.3	468.8	1.6	30	2,4,6,8
H2	24.19	250	1.6	32	1-8
H3	32	625	1.6	40	2,4,6,8
I1	39.5	10000	0.4	40	1,2

Table 4.2: Individual Wire Structure Characteristics

The parasitic values from Table 2 were entered into the multiplier module schematic in *Cadence*. Also, two 5 nH inductors modeling the bondwire were placed on the schematic. They were built using *ASITIC*. A high level schematic of how the

individual wire structures were entered in the multiplier module is shown in Figure 9. Please note that each horizontal impedance (H1-H3) and each interconnect impedance (I1) represents nine of the specified impedance in parallel. This was done to represent the different horizontal V_{dd} and G_{nd} planes shown in Figure 8. An example is shown in Figure 10.

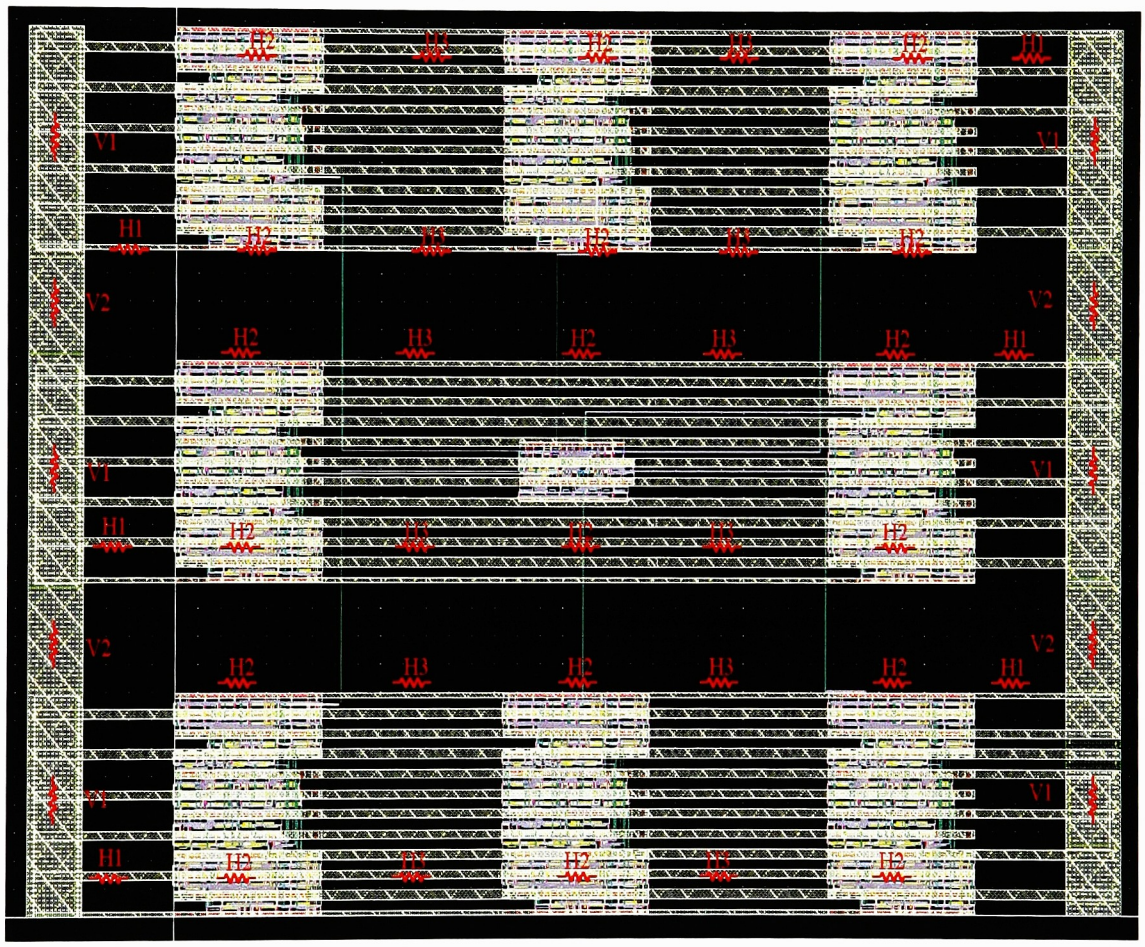

Figure 4.8: Interconnect Highlighted Multiplier Module

A simulation was run for 25 ns with an input clock of 2 GHz and the phase splitter was set so all output clocks were in phase. Figure 11 shows the difference between the supply voltage and ground voltage directly after the bondwire. This is an important metric because it represents the amount of voltage headroom the digital logic has to switch.

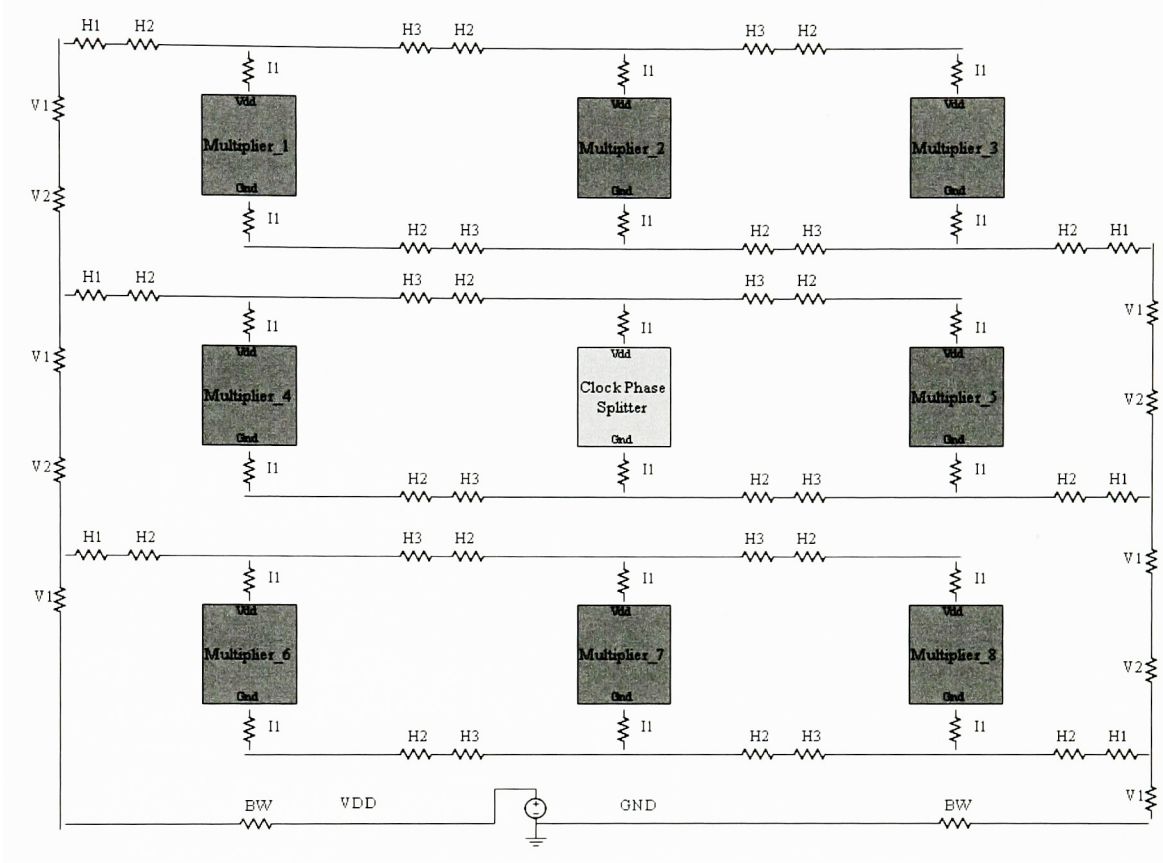

Figure 4.9: Multiplier Module Interconnect Model

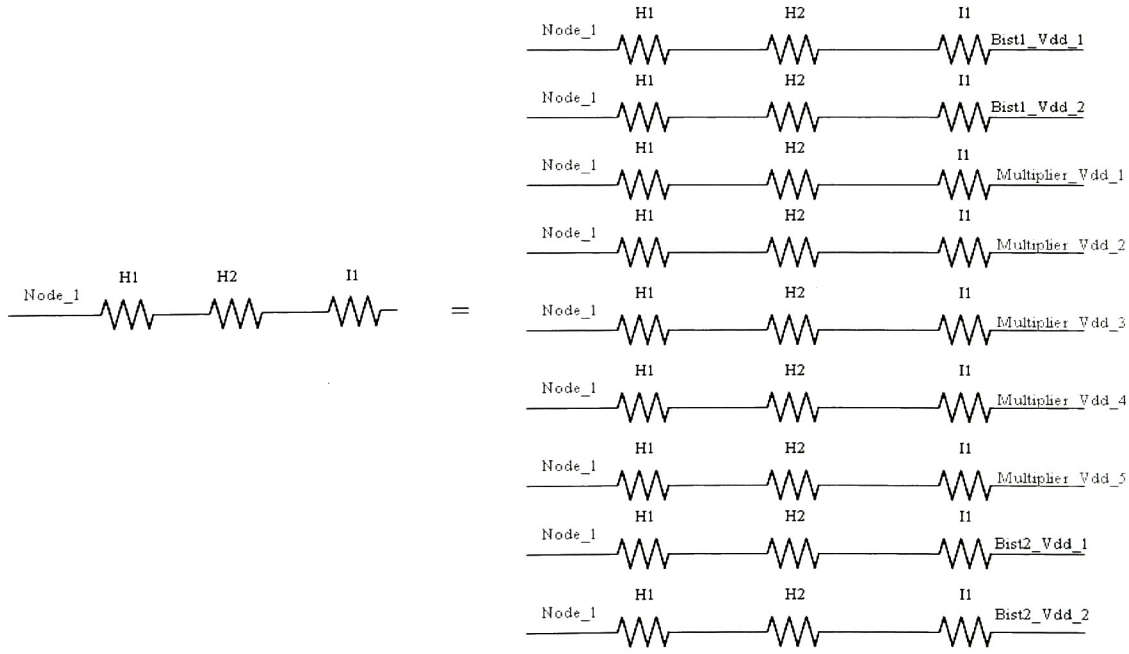

Figure 4.10: Impedance Representation

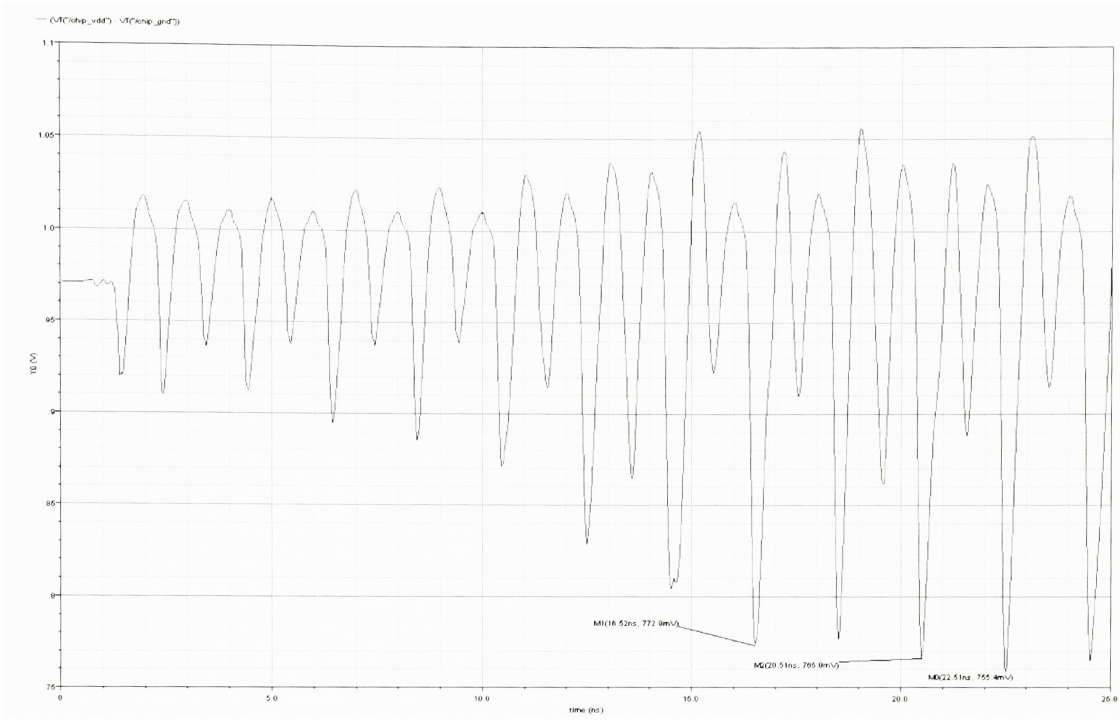

Figure 4.11: Difference Between Supply and Ground Voltage Bondwire

Figure 11 shows that the inductive bondwire has a significant impact on power supply stability. It was able to degrade the supply from 1 V to a worst case of 750 mV. This is well beyond targeted 50 mV of allowed fluctuations.

Voltage plots were also extracted on nodes directly before the multipliers, further into the interconnect. This was done to compare the difference between the bondwire headroom voltage and the interconnect headroom voltage. For example, if the voltage on the supply is 975 mV after bondwire and 960 mV on a node directly in front of a multiplier, the on-chip interconnect was responsible for a 15 mV drop.

Using Figure 9 to reference multiplier 2; the supply voltage and ground voltage were measured directly after the I1 impedance line. Multiplier 2 was chosen because it has the most impedance between it and the supply. The headroom voltage is shown in Figure 12.

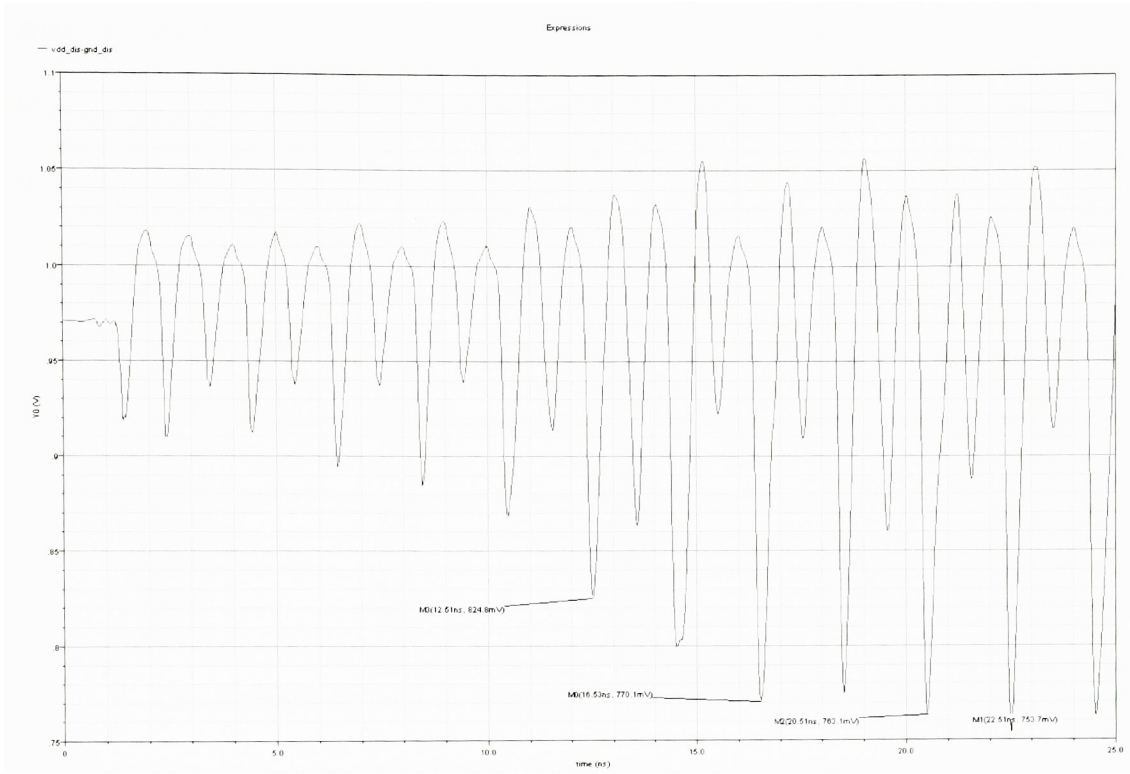

Figure 4.12: Difference Between Supply and Ground Voltage Multiplier 3

Figure 12 shows a worst case headroom voltage of 748 mV. This implies that the interconnect structure was responsible for an extra 7 mV of supply noise after the bondwire.

Table 3 shows the worst case amount of headroom voltage for each multiplier block. The difference between each multiplier block headroom voltage and the bondwire headroom voltage is also listed. The worst case is 7 mV in multiplier block 2. Even though the digital structure simulated in this work did not generate enough noise to degrade the noise margin, it does provide a basis for examining realistic digital structures. If the results shown in Table 3 are extrapolated to model a digital circuit with millions of transistor, it can easily be seen how important an effective decoupling strategy would be. For example, the amount of interconnect generated supply noise on a processor would be much higher than what was shown in this work and would require serious attention.

Also, as supply voltages continue to shrink, the amount of tolerable noise will also reduce. A 15 mV interconnect drop on a 500 mV supply may have much worse effects than the same drop on a 1 V supply.

Multiplier Block	Worst Case BW Headroom (mV)	Worst Case Headroom (mV)	Difference (mV)
Block 1	755	752	3
Block 2	755	748	7
Block 3	755	750	5
Block 4	755	749	6
Block 5	755	750	5
Block 6	755	752	3
Block 7	755	750	5
Block 8	755	752	3

Table 4.3: Multiplier Noise Comparison

The analysis derived in Chapter 3 was used to place decoupling capacitance within each multiplier module layout. Four different strategies were implemented and an activity factor of 5% was used. The first was a lumped strategy, the middle two were grouped strategies and the last was a distributed strategy. The process used to calculate decoupling capacitance implemented the ideas presented in Chapter 3 (Equations (6) to (20)). The basic equation utilized was (3.16):

$$C_{decap} = \frac{Q_{node}}{V_{dd}\ \lambda_{node}}$$

$$C_{decap} = \frac{Q_{node}}{V_{dd}\left(1 - \frac{V_{dd\,\min}}{V_{line}}\right)}$$

For example, referencing Figure 9, if a decoupling capacitor was going to be placed in front of multiplier block 2 the amount of high frequency switching charge drawn by multiplier 2, Q_{node}, was found via simulation. An example plot of one of the supply nodes within multiplier block 2 is shown in Figure 13. There are 9 supply nodes

within multiplier block 2. These can be seen on the layout in Figure 7. The amount of charge for each node is calculated and summed to determine the amount of charge capable of being switched by the multiplier block.

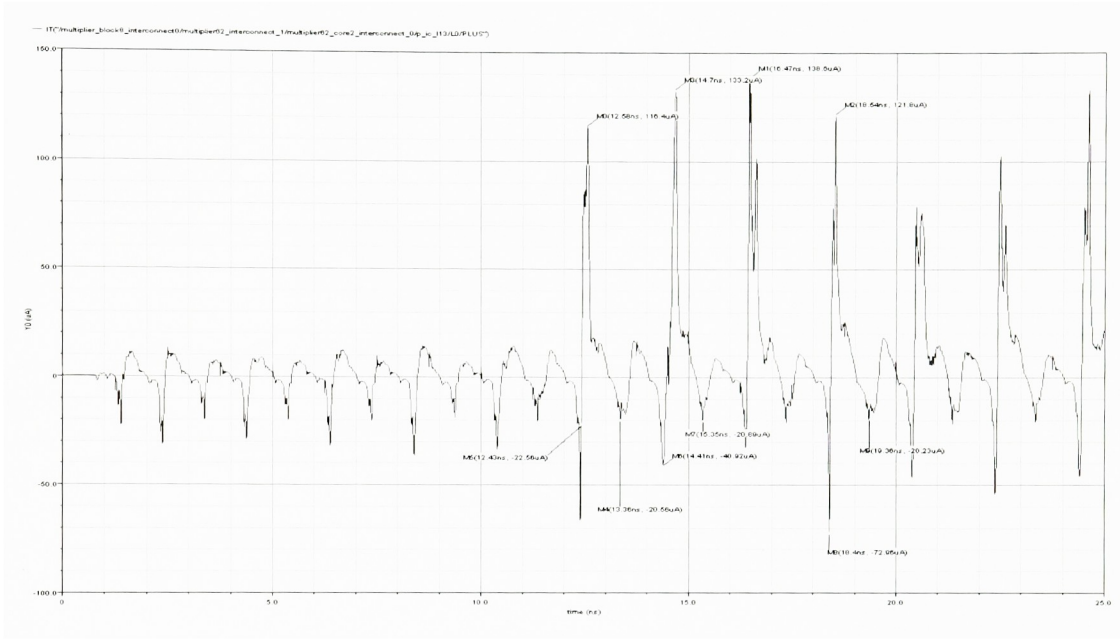

Figure 4.13: Multiplier 3 High Frequency Switching Charge

$V_{dd\,\min}$ for the given activity factor was 950 mV. V_{line} was gotten from the amount of resistance drop seen on the line for multiplier 3 before any switching activity occurred. This is shown in Figure 14.

These values were then substituted into Equation (3.17) to solve for decoupling capacitance. For example, in the above figures, Q_{node} is 72 fC, and V_{line} is 967 mV. This leads to 4.096 pF of decoupling capacitance.

MOS capacitance was used to physically layout the decoupling capacitors. This structure consists of a single NMOS transistor with the gate connection tied to the supply and the source and drain connections tied to ground. The capacitance formula used was:

$$C_{MOS} = WLC_{OX}$$
$$W = 2L$$
$$C_{MOS} = 2L^2 C_{OX}$$

Device width was twice device length to increase the amount of series resistance between the capacitor and ground (help dampen the effects of oscillations). Using the above example:

$$C_{decap} = \frac{72\,fC}{1V\left(1 - \frac{950\,mV}{967\,mV}\right)} = 4.096\,pF$$
$$C_{MOS} = C_{DECAP} = 2L^2 C_{OX}$$
$$C_{OX} = \frac{\varepsilon\varepsilon_0}{t_{ox}} = \frac{3.9(8.854x10^{-14}\,\frac{F}{cm})}{72x10^{-8}\,cm} = 4.796x10^{-7}\,\frac{F}{cm^2}$$
$$L = 20.66\,um$$
$$W = 41.33\,um$$

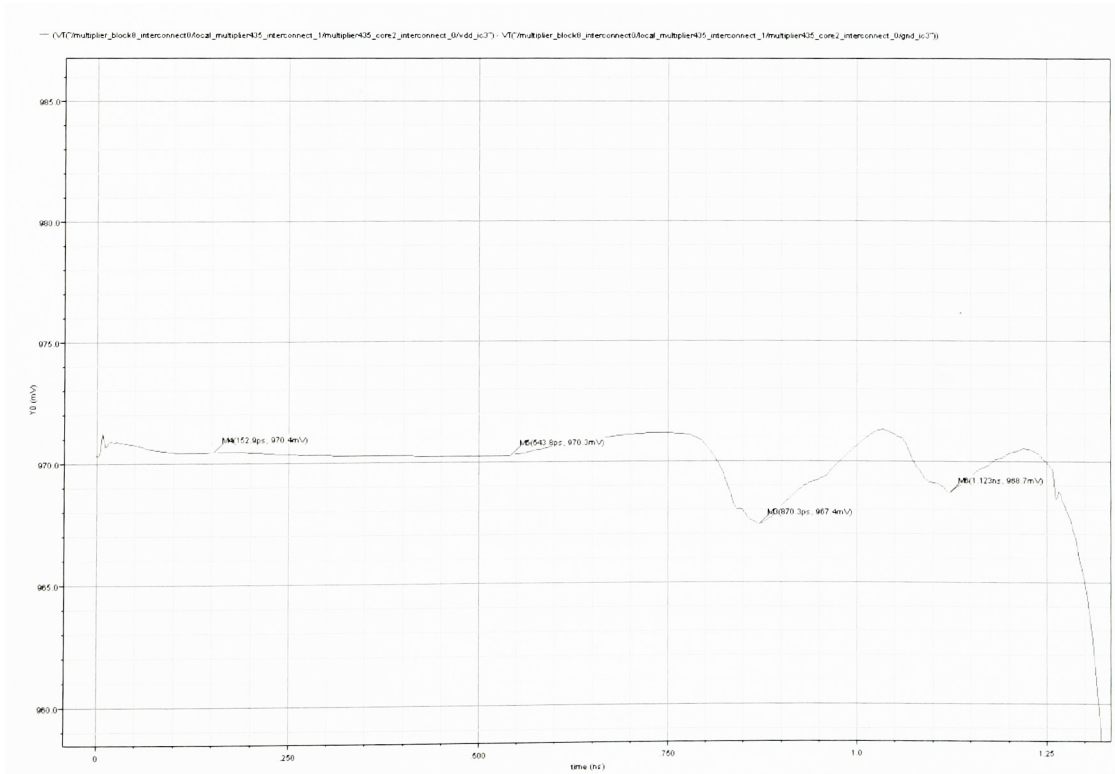

Figure 4.14: VLINE Calculation

Figure 15 shows the physical layout of a decoupling capacitor.

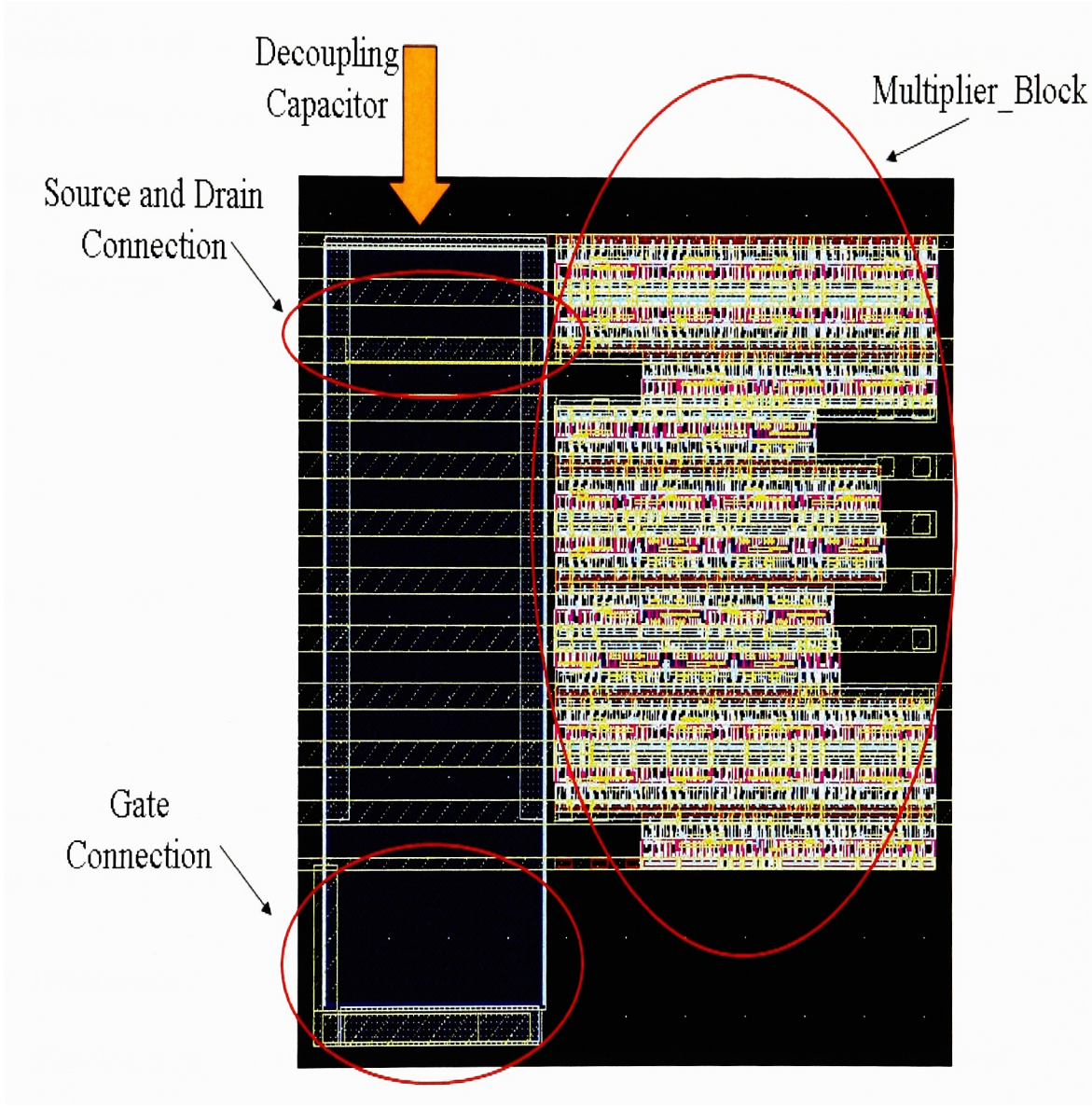

Figure 4.15: Decoupling Capacitor Physical Layout

The following three sub-chapters present the four different decoupling strategies. The procedure explained above was used to determine the amount of capacitance to be placed for each different strategy.

4.3.1 Lumped

The lumped strategy was followed in such a manner that two capacitors, each approximately 16 pF, were used in a single multiplier module. The layout is shown in Figure 16. These two capacitors are responsible for providing high frequency charge to each multiplier.

4.3.2 Grouped I

The first grouped strategy placed one decoupling capacitor per multiplier. Each capacitor was valued at 7.5 pF. The layout is shown in Figure 17. This is a grouped approach in the sense that a decoupling capacitor was placed in front of each multiplier.

4.3.3 Grouped II

The second grouped strategy placed 8 decoupling capacitors per multiplier. Each capacitor was valued at 0.673 pF. The layout is shown in Figure 18. Six decoupling capacitors were used per multiplier. A single decoupling was placed for every horizontal supply wire and ground wire in a multiplier.

4.3.4 Distributed

The distributed strategy placed 16 decoupling capacitors per multiplier. Each capacitor was valued at 0.389 pF. The layout is shown in Figure 19. Eleven decoupling capacitors were used per multiplier. A single decoupling capacitor was placed per horizontal row of primitive elements within the multiplier.

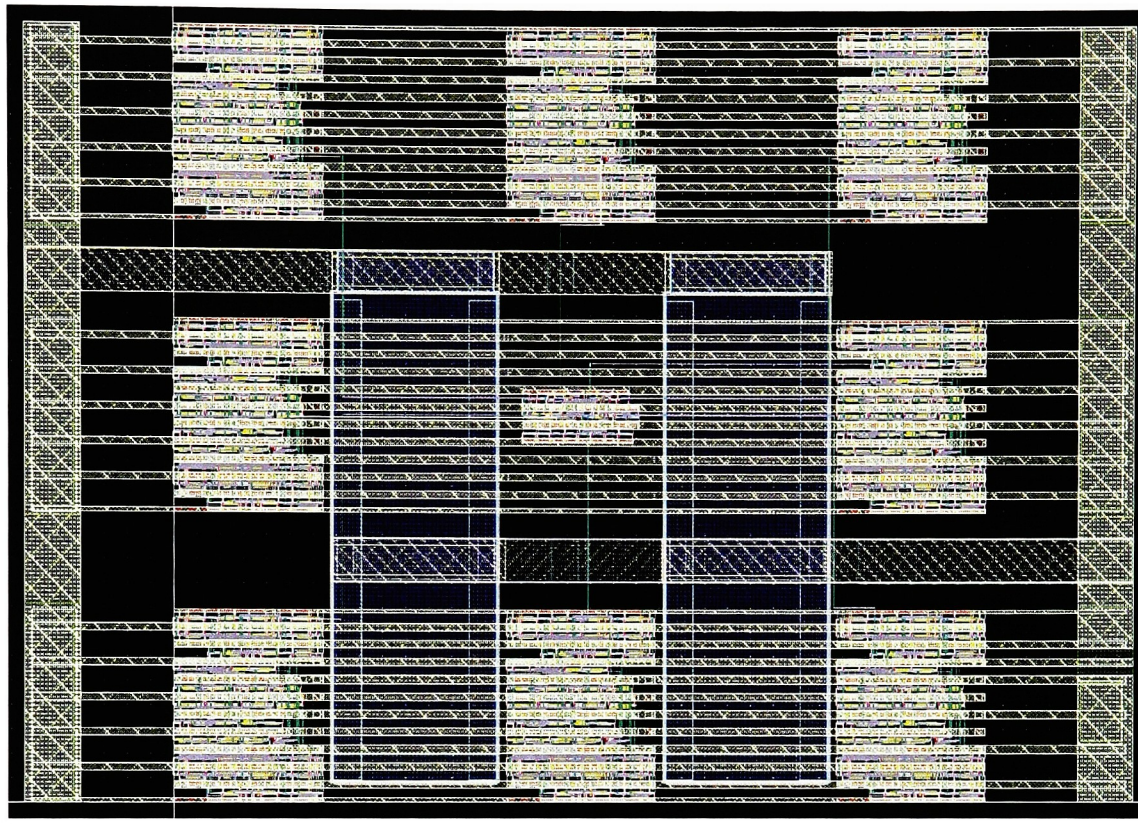

Figure 4.16: Lumped Decoupling Strategy

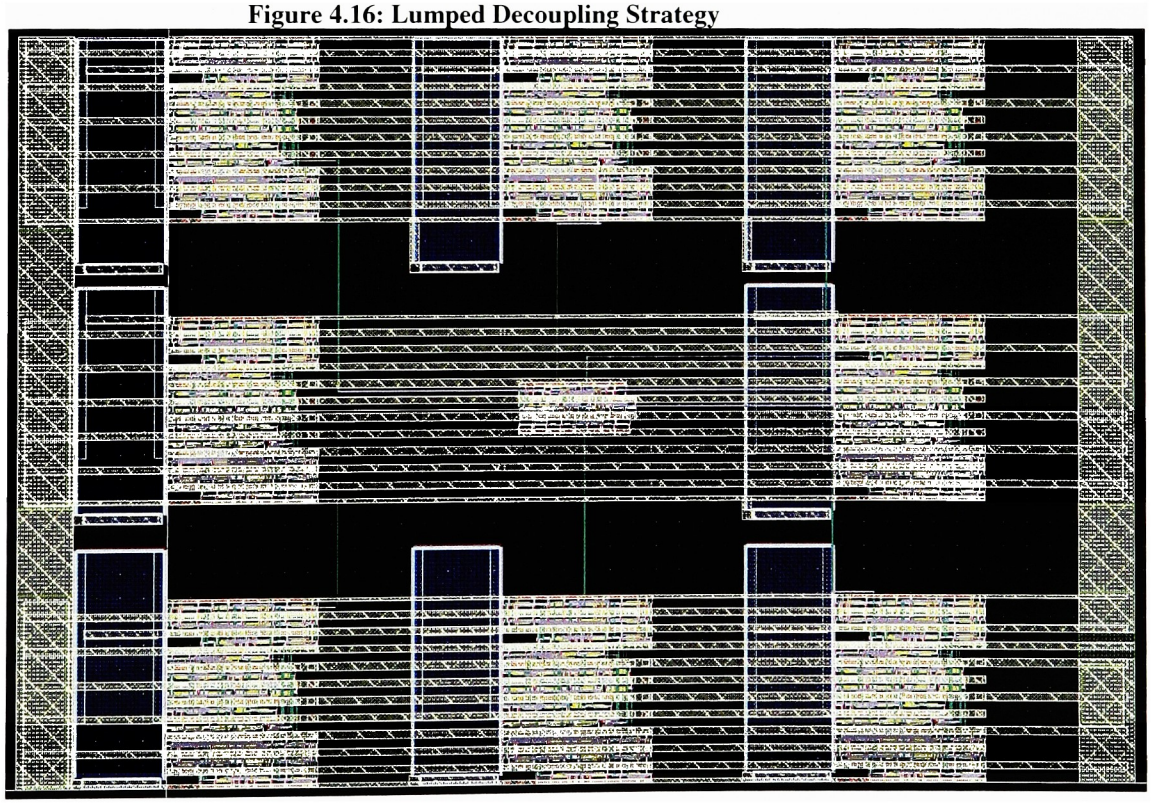

Figure 4.17: Grouped I Strategy

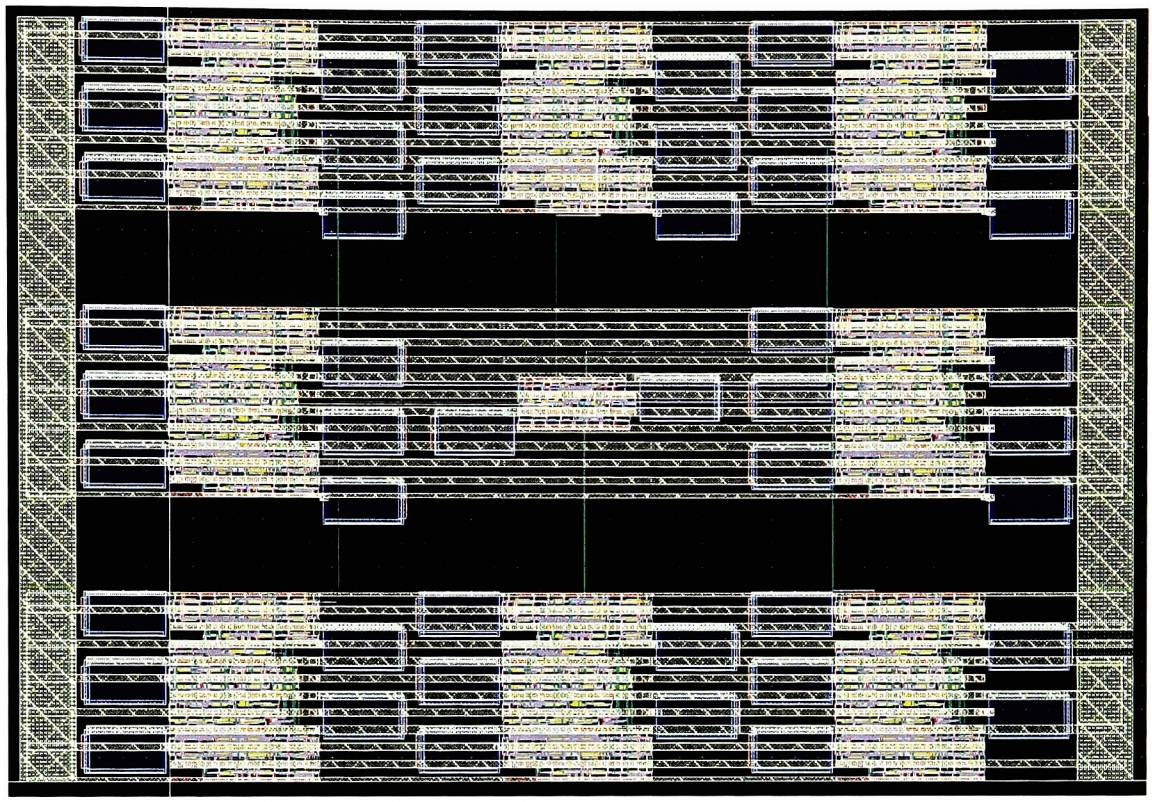

Figure 4.18: Grouped II Strategy

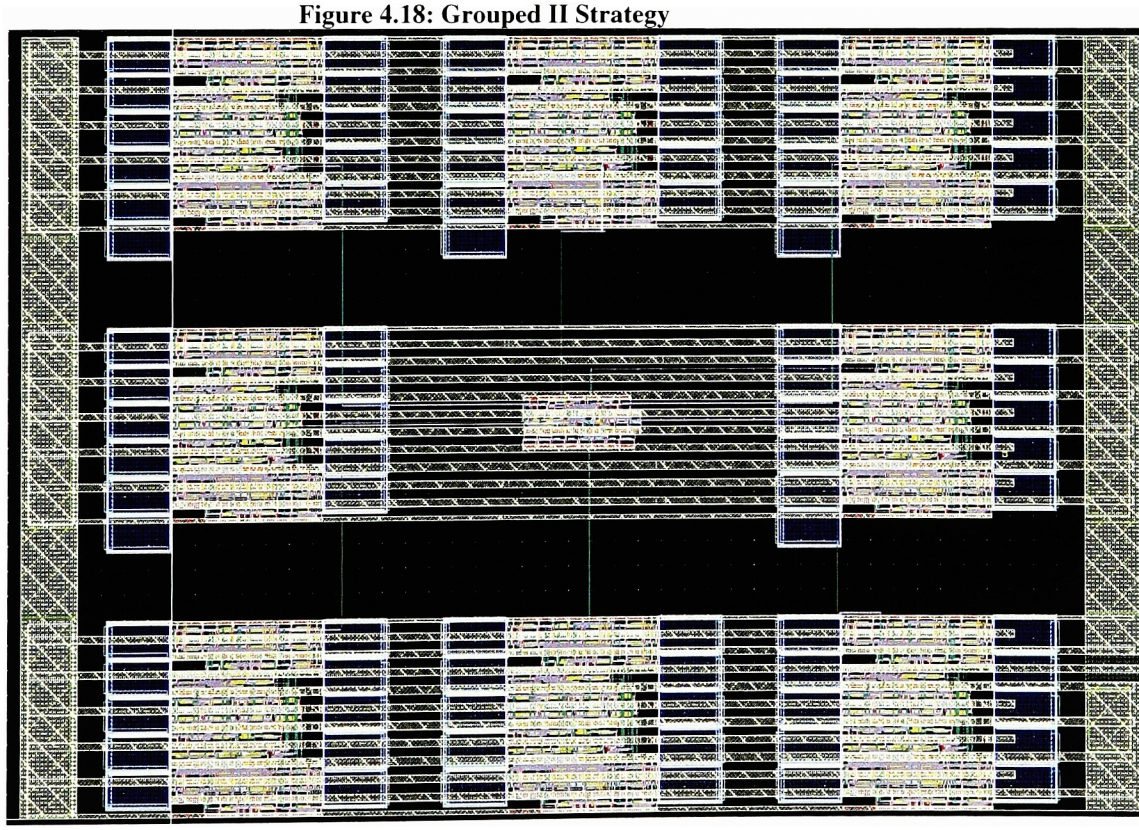

Figure 4.19: Distributed Strategy

Decoupling Strategy	Single Capacitor Value (pF)	Width/Length (μm)	Number of Single Capacitors	Total amount of capacitance (pF)	Total amount of area (μm^2)
Lumped	16.786	100/35	2	33.572	6700
Grouped I	4.2	47/18	8	32.6	6768
Grouped II	0.673	16.5/8.5	50	33.65	7012.5
Distributed	0.389	12.5/6.5	88	34.232	7150

Table 4.4: Decoupling Capacitance Comparison

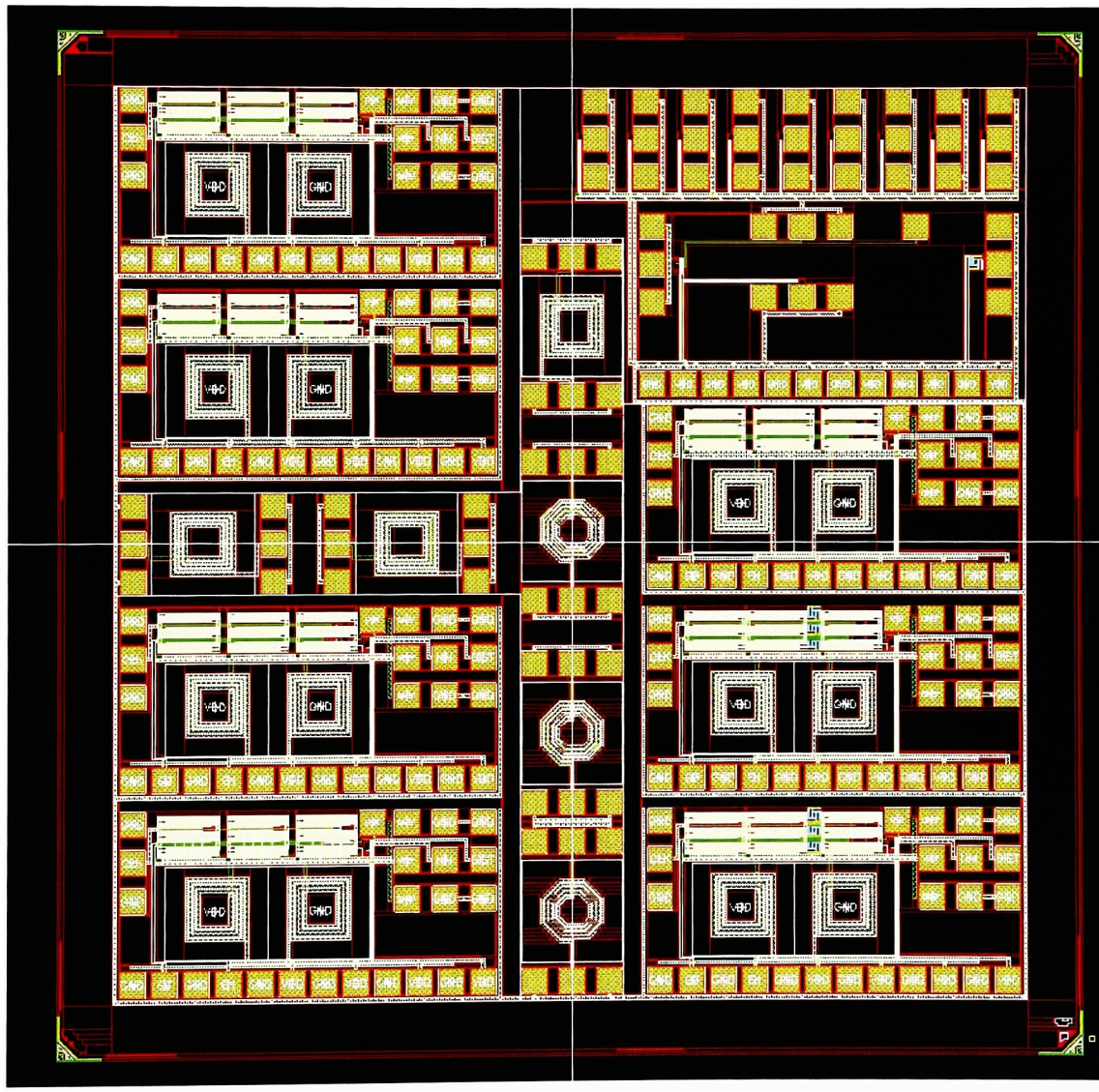

Figure 4.20: Complete Test Chip

Table 4 shows a comparison of all decoupling strategies mentioned above. The lumped strategy took the least amount of area while the distributed strategy had the

highest amount of capacitance per area. One of the grouped strategies may be optimal in the sense that it combines a relatively low amount of area consumption with a high amount of capacitance per area. The complete test chip is shown in Figure 20. The different decoupling strategies are implemented on the left hand side of the chip. The right hand side of the chip contains the circuit presented in [30].

4.4. Noise Monitoring Techniques

In order to compare the different decoupling strategies, there must be some way to monitor the amount of on-chip noise. There are two elements to noise monitoring that must be taken into account. The first is where to measure the supply noise within the interconnect structure and the second is the actually circuitry that will do the monitoring.

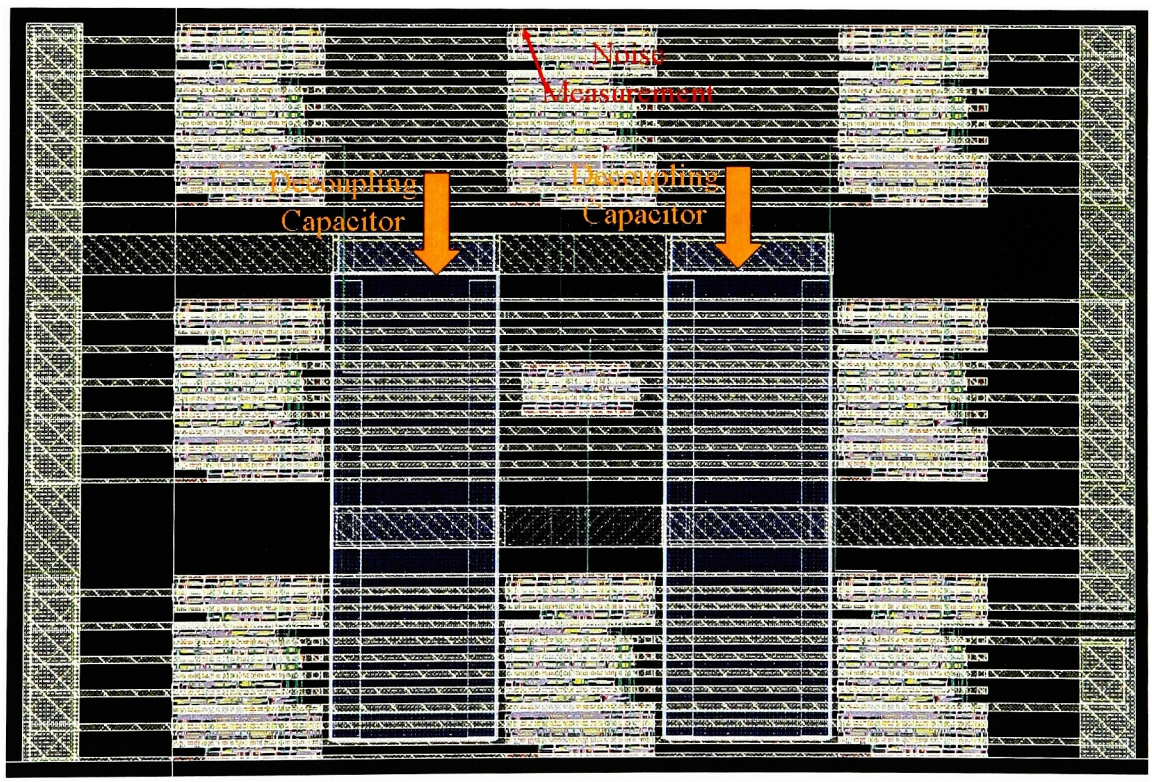

Figure 4.21: Noise Measurement Location

In order to determine if one decoupling scheme has advantages over another, the amount of supply noise generated must be compared. The noise measurement should take place as close to the switching elements as possible because that is where the stability of the supply voltage is most important. For example, if a lumped decoupling scheme was used within an interconnect structure to reduce noise then the amount of supply noise should not be measured where the lumped capacitor was placed. The noise should be measured further in the interconnect structure. This idea is shown in Figure 21.

Two possible techniques have been targeted as a way to monitor noise and thus compare different decoupling strategies. The first is a spectrum analyzer, which can monitor signals with frequencies up to 4 GHz. Depending on how much noise is present on-chip, the spectrum analyzer may not be sensitive enough to detect small fluctuations in noise.

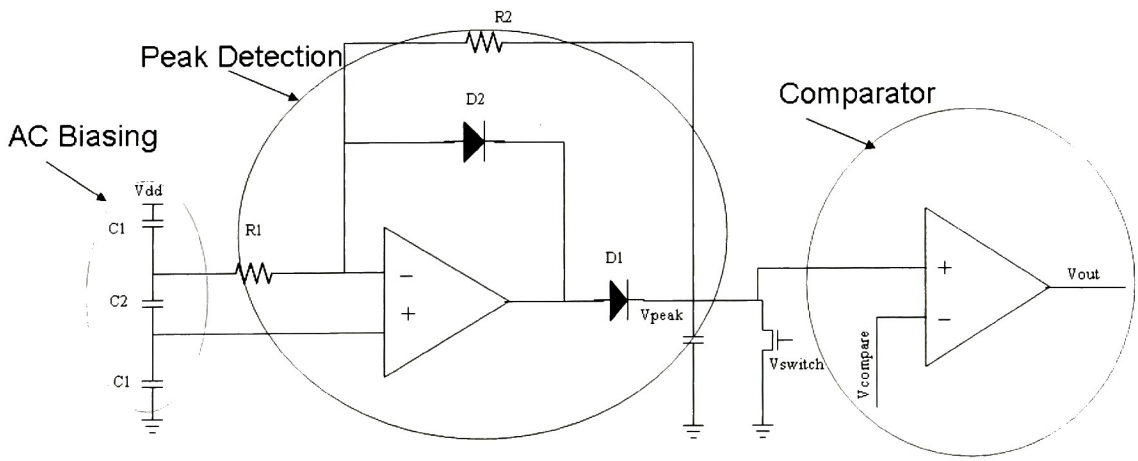

Figure 4.22: Noise Monitoring Schematic

The second technique is to make use of the circuit presented in [30], where a differential amplifier is used to detect the noise between the supply and ground. In [30], the detected signal is injected into ground through an AC coupling capacitor in hopes of

suppressing noise seen on the supply. This approach would have to be modified to monitor noise and not suppress noise. A high level schematic is shown in Figure 22. It should be noted that this circuit was not implemented. This schematic uses the same capacitor network as presented in [30] to detect supply noise and feeds that to a precision rectifier circuit. The precision rectifier works as a peak detector and could be used to detect the peak values seen on the power supply. These peak values are then sent to a comparator and compared to a reference voltage. If the peak value on the supply suddenly drops lower than the reference voltage, the comparator would switch signifying noise on the supply.

Chapter 5 Discussion

5.1. Optimal Decoupling Capacitance Strategy

Intuitively, the high frequency current supplied by a decoupling capacitor should encounter as little resistance and inductance as possible. If this happens, voltage drops are limited and the supply is kept stable. This idea is depicted in Figure 1

Figure 5.1: Decoupling Capacitor Comparison

The easiest way to accomplish this is to place a capacitor in each primitive cell. For example, in a digital design, each inverter would have its own decoupling capacitor within its respective layout. This idea would also apply to other gates like NAND, NOR, and XNOR. Simulations or complex gate collapse techniques could be used to determine the amount of charge needed to be supplied by each decoupling capacitor. With regards to the three generic decoupling strategies introduced in chapter 3, placing a decoupling capacitor into each primitive cell is a completely distributed approach. Capacitance cannot be placed any deeper into the interconnect than directly in front of each cell.

This approach may not be practical because of physical limitations. Small capacitance values, possible leakage currents, and area overheads may create a need to group primitive cells into larger blocks and place decoupling capacitance in front of each group. This follows more of a grouped approach because a single decoupling capacitor would be used to provide charge to multiple switching cells.

If a true optimal solution is going to be found, the designer must have some control over the smallest value of decoupling capacitance allowed. This would help prevent some of the previously mentioned physical problems from happening. Assuming that a fully distributed approach is the best way to create a stable power supply because the charge supplied by the capacitors encounter as little impedance as possible, one optimal solution may be starting from a completely distributed approach and increasing the level of distribution until the decoupling capacitance values are greater than a limit set by the designer. Obviously, the second order parameters shown on Chapter 3 would have to be analyzed for a solution to be accepted.

Another approach to finding an optimal placement strategy is to increase the level of distribution from lumped to distributed and find where improvements on supply noise begin to become minimal. For example, if a level 3 distribution shows 20% noise improvement over a level 2 distribution and a level 4 distribution shows only a 2% noise improvement over level 3 distribution, the level 3 distribution could be called optimal. Again, the second order parameters shown in Chapter 3 would have to be analyzed for strategy acceptance.

5.2. Practicality

If the ideas presented in this work are going to be implemented in a real chip design, the entire process would need to be automated. It is unrealistic to think a designer of a large digital chip would take the time to go through this analysis after completing a layout. Automating the placement of decoupling capacitance allows the designer to focus on other more important aspects of physical implementation like functionality and structural integrity.

If automation were to occur, the theory in chapter 3 would have to be programmed into a CAD tool like *Cadence*. Once the layout of the circuit is complete, the interconnect structure would be extracted. It is important that the extraction include parasitic resistance, inductance, and capacitance. In some versions of *Cadence*, extracted schematics only include resistance and capacitance. Because this work focuses on inductance as a primary power supply noise source, it is necessary to be included within the extraction. Inductance values change at different frequencies, so techniques like those used in *ASITIC* would have to be implemented.

Once the interconnect structure has been formed, the parasitic values can be inserted back into the original schematic. Within the CAD software, a decoupling capacitance analysis could be run for different levels of distribution. Once this analysis is complete, the user would have the ability to select the desired type of decoupling scheme. Obviously, there would need to be user defined parameters for the automated process to be complete. The amount of work needed to automate this process would be extensive and time consuming but it would help ease some responsibilities of the designer.

Chapter 6 Conclusion

6.1. Observations

It was the intent of this work to provide an on-chip methodology to locate and quantify decoupling capacitance. The motivation for performing this work derived from the relationship between shrinking transistors, power supply noise, and the on-chip interconnect structure. Often, designers do not take decoupling capacitance into account until layout is complete, when they randomly fill empty space with capacitance. This work provided a methodology to allow designers to consider decoupling capacitance while they perform layout instead of an annoying task that needs to be done after layout.

Previous research involving decoupling capacitance looked at quantification or location, not both. This work analyzed both aspects at the same time. The results showed that the most effective way to reduce supply noise is to distribute capacitance as close to the switching elements as possible. If this type of distributed approach is taken, there are certain trade-offs that must be accounted for; layout complexity, delay times, decreased natural frequency, and increased capacitance. This work derived a complete mathematical model that attempted to calculate how much decoupling capacitance to place within an interconnect structure. Also, metrics were derived to be used as a guide to locate decoupling capacitance.

This work demonstrates a proof of concept; how quantification and placement of decoupling capacitors can be performed on generic interconnect structures. Although the digital logic simulated is small with regards to area, the results provide an initial basis for determining how detrimental supply noise can become within large interconnect

structures. If the results are extrapolated to model a large digital circuit, it is easily understood how important decoupling capacitance placement is with a 1 V supply. Although transistor noise margins are relatively high in the TSMC 90 nm process, on-chip devices must be able to handle noise from multiple different sources. Because these different sources often generate noise simultaneously, it is imperative that effective solutions exist for each of them. The importance of quantification and placement of decoupling capacitance will continue to rise as noise margins shrink. This will happen as supply levels decrease, which according to Figure 1 in Chapter 1, will happen in the near future.

6.2. Future Work

Improvements upon this work could be made by modeling the on-chip interconnect structure in a more accurate manner. The structure was modeled as a tree between the supply and switching elements. In actuality, an interconnect structure would also exist between the switching elements and ground, as is shown in Figure 2. This would lead to more noise on the supply in the sense that voltage spikes would occur on both the high potential and lower potential interconnect lines. Modifications of the theory presented in Chapter 3 would have to be made because of the added resistance and inductance, which would be placed after the current sources within the RLC tree.

Also, the current model used for switching was a triangular pulse. While effective, this model may be too simplistic and could be a source of errors in predicting how high frequency switching flows. A more accurate model that contains predictable parameters could be used to characterize how current is being pulsed on-chip. The ability to accurately model current draw is directly related to the effectiveness of the decoupling

capacitors. The amount of charge moved during a clock pulse is derived from the current draw and used to solve for the decoupling capacitance. It is understood that in a complex digital circuit, current prediction can be somewhat of a mystery because different components will be switching at various times. Given this constraint, a robust model needs to be developed that can predict nominal current draw for digital circuits.

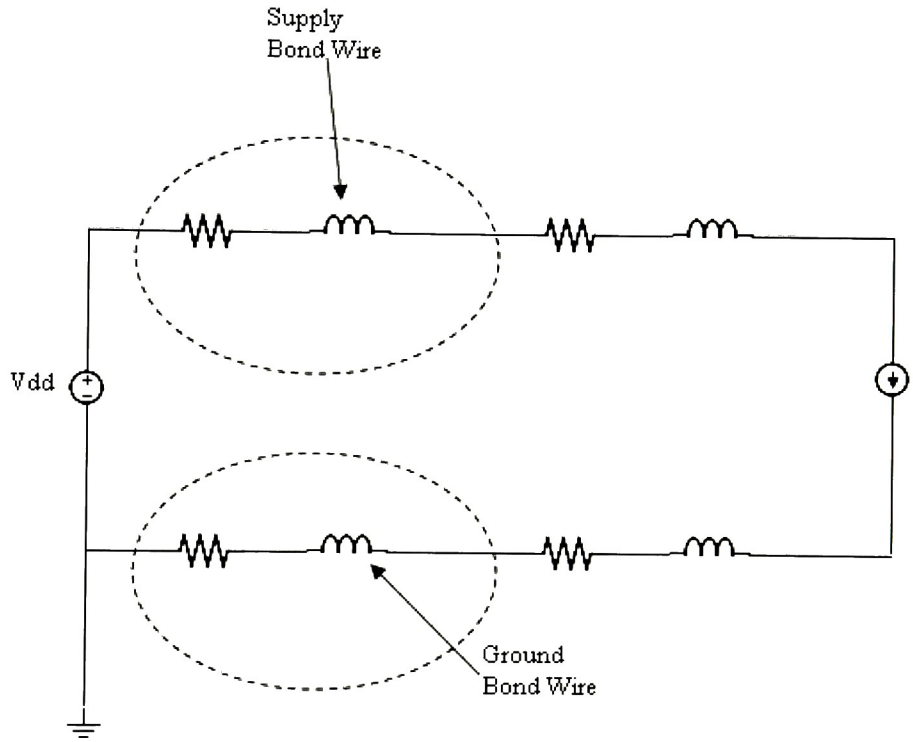

Figure 6.1: Distribution Network with Ground Impedance

The metrics developed in Chapter 3 are only reliable for a binary tree. Modifications must be made if they are going to applicable to realistic interconnect structures.

The type of decoupling capacitor to be used in physical layout is another important consideration. Different types of structures can be used as decoupling capacitance. These include MOS capacitances, PIP, capacitances, MIM capacitances, BR capacitances, and other structures that may be capable of producing a reliable amount of

capacitance. Each of these different structures has advantages and disadvantages. For example, a thin oxide MOS capacitor may produce a high amount of leakage current if sized too small. A MIM capacitor may require a large area overhead to meet a targeted, frequency independent value. A comparison of these structures being used as decoupling capacitors would be useful to help determine which type of capacitor to use in different situations.

The ability to monitor supply noise is equally important as rejecting it. If noise can't be accurately monitored, there will be no way to determine how effective different strategies are at rejecting it. Some type of reliable circuitry that can detect high frequency noise is needed to validate this work.

Bibliography

[1] Moore, G.E., "Cramming More Components Onto Integrated Circuits," Proceedings of the IEEE , vol.86, no.1, pp.82-85, Jan 1998.

[2] L. D. Smith, R. E. Anderson, D. W. Forehand, T. J. Pelc, and T. Roy, "Power distribution system design methodology and capacitor selection for modern CMOS technology," *IEEE Trans. Adv. Packag.*, vol. 22, p. 284, Aug. 1999.

[3] John P. Uyemura. Introduction to VLSI Circuits and Systems. Wiley and Sons, 2002.

[4] C. T. Tsai, "Package inductance characterization at high frequencies," IEEE Trans. Comp., Packag., Manufact. Technol., part B, vol. 17, pp. 175-181, May 1994.

[5] M. Pude, "An Analytical Propagation Delay Model Including Power Supply Noise," Graduate Paper, Department of Electrical Engineering, Rochester Institute of Technology, November 2005.

[6] M. D. Pant, P. Pant, and D. S. Wills, "On-chip decoupling capacitor optimization using architectural level prediction," IEEE Trans. on VLSI Syst., Vol. 10, pp. 319–326, June 2002.

[7] R. Downing, P. Gebler, and G. Katopis, "Decoupling capacitor effects on switching noise," in *IEEE Topical Meeting Elect. Performance Electron. Packaging*, 1992, pp. 148–151.

[8] Q.K. Zhu, "Power Distribution Network Design for VLSI", John Wiley & Sons, 2004, Chapter 1.

[9] Y. I. Ismail, E. G. Friedman, and J. L. Neves, "Equivalent Elmore Delay for RLC Trees," in Proc. IEEE/ACM Design Automation Conf., June 1999, pp. 715-720.

[10] Y. I. Ismail, E. G. Friedman, and J. L. Neves, "Signal Waveform Characterization in RLC Trees," Circuits and Systems, July 1999. ISCAS '99.

[11] Y. I. Ismail, and E. G. Friedman, "Inductance Effects in RLC Trees", VLSI, 1999. Proceedings. Ninth Great Lakes Symposium on 4-6 March 1999 Page(s):56 – 59.

[12] S. Zhao, K. Roy, and C.-K. Koh, "Decoupling Capacitance Allocation for Power Supply Noise Suppression," In Proc. International Symposium on Physical Design, pages 66-71, Napa, CA, April 2001.

[13] M. Popovich and E. G. Friedman, "Maximum Effective Distance of On-Chip Decoupling Capacitors in Power Distribution Grids," 16th ACM Great Lakes Symposium on VLSI, April 30 – May 2, 2006.

[14] K.Wang and M. Marek-Sadowska, "On-chip power supply network optimization using multigrid-based technique," in *IEEE Design Automation Conf.*, 2003, pp. 113–118.

[15] H. Chen and J. S. Neely, "Interconnect and circuit modeling techniques for full-chip power supply noise analysis," *IEEE Trans. Comp., Packag., Manufact. Technol. B*, vol. 21, pp. 209–215, Aug. 1998.

[16]M. Sotman, A. Kolodny, M. Popovich, and E. G. Friedman, "On-die Decoupling Capacitance: Frequency Domain Analysis of Activity Radius," ISCAS 2006.

[17] C. A. Lin, and C.-H. Wu, "Second-Order Approximations for RLC Trees," IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, Vol. 24, No. 7, July 2004.

[18] T. Sakurai, "Approximation of wiring delay in MOSFET LSI," *IEEE J. Solid-State Circuits*, vol. SC-18, pp. 418-426, Aug. 1983.

[19] A. Deutsch , et al., "Modeling and characterizing of long interconnections for high-performance microprocessors," *IBM J. Res. Develop.*, vol. 39, no. 5, pp. 547-667, Sept. 1995.

[20] H.H. Chen, J.S. Neely, M.F. Wang, G. Co, "On-chip Decoupling Capacitor Optimization for Noise and Leakage Reduction", Proceedings of 16th Symposium on Integrated Circuits and Systems Design, SBCCI'03. pp. 251-255.

[21] L. D. Smith, "Decoupling capacitor calculations for CMOS circuits," in Proc. Elect. Perf. Elect. Packag. Conf., Nov. 1994.

[22] M. Popovich, and E. G. Friedman, "Decoupling Capacitors for Multi-Voltage Power Distribution Systems," IEEE Trans. on VLSI Syst., Vol. 14, No. 3, March 2006.

[23] H. Su, S. Sapatnekar and S. Nassif, "Optimal decoupling capacitor sizing and placement for standard-cell layout designs," *IEEE Trans. on CAD*, vol. 22, no. 4, April 2003.

[24] J. Fan, J. L. Knighten, A. Orlandi, N. W. Smith, and J. L. Drewniak, "Quantifying decoupling capacitor location," presented at the *Proc. IEEE Int. Symp. Electromagnetic Compatibility*,Washington, DC, Aug 2000.

[25] P. Larsson, "Resonance and damping in CMOS circuits with on-chip decoupling capacitance," IEEE Trans. Circuits Syst. I, vol. 45, pp. 849–858, Aug. 1998.

[26] P. Larsson, "Parasitic resistance in an MOS transistor used as onchip decoupling capacitor," *IEEE J. Solid-State Circuits*, vol. 32, pp. 574–576, Apr. 1997.

[27] J. Choi, S. Chun, N. Na, M. Swaminathan, and L. Smith, "A Methodology for the Placement and Optimization of Decoupling Capacitors for Gigahertz Systems," Presented at the International Conference on VLSI Design, Calcutta, India, Jan. 2000.

[28] C. R. Paul, "Effectiveness of multiple decoupling capacitors," *IEEE Trans. Electromagn. Compat.*, vol. 34, pp. 130–133, May 1992.

[29] A. Niknejad, "Modeling of passive elements with *ASITIC*," in *Proc. IEEE RFIC Conf.*, June 2002, pp. 303–306.

[30] C. Urban, "Active Noise Cancellation Using Feed Forward Techniques," Masters Thesis, Rochester Institute of Technology, to be published.